Department of Computer Science,
University of Otago

Technical Report OUCS-2017-02

An extended model of deictic routines, supporting a wider-coverage SM interpretation of syntax

Author:

Alistair Knott
Department of Computer Science, University of Otago, New Zealand

Department of Computer Science,
University of Otago, PO Box 56, Dunedin, Otago, New Zealand

http://www.cs.otago.ac.nz/research/techreports.php

An extended model of deictic routines, supporting a wider-coverage SM interpretation of syntax

Ali

August 29, 2016

Contents

Chapter 1

Introduction

In this chapter I'll summarise the idea that the LF of concrete sentences can be interpreted in SM terms. Basically give a summary of Knott (2012).

Then I'll outline the structure of the rest of the book.

Part I

Spatial representations, spatial actions and prepositional phrases

This part of the book presents a model of the cognitive system that perceives and represents environments and places, and of how operations within this system are represented in language. It contains four chapters. Chapter **??** introduces a model of the brain's spatial representations. Chapter **??** introduces a model of actions that update spatial representations in various different ways. Chapter **??** introduces a model of the internal syntax of prepositional phrases (PPs). Chapter 7 presents a sensorimotor interpretation of the internal syntax of PPs.

Chapter 2

A model of the brain's spatial representation system

2.1 Overview

A key idea in the model I will present is that the cognitive system that represents environments and places is quite distinct from, though of course linked to, the system that represents objects 'semantically' using a set of open-class types. I will call the former system the **environment/place system** and the latter system the **object system**. I will argue that while operations in the object system are described linguistically by DPs, operations in the environment/place system are described by PPs. For instance, a PP like *in the kitchen* describes a sensorimotor process in the environment/place system in which a place is attended to within a certain environment (the kitchen); the DP *the kitchen* within this structure refers to an object at this place.

In my proposal, space is represented in two systems, working together: one for representing environments, and one for representing places within environments. Each environment representation is associated with a map of places, whose topography relates to the shape of that environment. At some places in a given environment, neighbouring environments or sub-environments can be accessed: so the mechanism for transitioning between environments involves both environment representations and place representations.

My proposal is that the environment/place system is relatively autonomous, and can work quite independently from the system that represents objects in semantically sophisticated ways. By itself, the environment/place system allows an agent to navigate through a complex three-dimensional world, to locate and remember interesting places in this world, and to perceive other objects or agents moving through this world.

Like the other cognitive systems I have considered, I suggest that the environment/place system operates with deictic representations, and that the operations it carries out can be modelled as deictic routines. A key concept for the model is that the environment/place system is always *referred* to a particular entity, which I will call its 'subject': this can be the observer himself, or an external entity that the observer is perceiving or remembering.

The environment/place system represents environments and places relating in some way to this subject—in other words, environments and places that are deictically referred to this subject. There are different ways in which they can relate to the subject: I will suggest that these relate to the different syntactic environments in which PPs can occur. But they always refer to a subject: there is no such thing as a 'stand-alone' representation within the environment/place system.

The environment/place system has two main components. One component is sensorimotor. There are perceptual mechanisms for representing the environment the observer is in, and his place within this environment, or other places which are observed. There are also perceptual mechanisms for representing other environments the observer can perceive, but is not currently in. And there are motor mechanisms for moving through the world, and for representing the movements of other entities. The other component relates to memory—specifically, to long-term memory (LTM). While the object and episode representation systems discussed earlier also have sensorimotor and memory components, the memory components I discussed in these cases are mainly to do with working memory (WM). In the case of the environment/place system, the memory component involves LTM representations very prominently, alongside WM representations. A distinguishing feature of the environment/place system is that it stores information about the spatial structure 'of the whole world': clearly this is held in LTM rather than WM. In the model I propose, the LTM representation of the world takes the form of a graph, in which (at a first approximation) the nodes represent environments, and the arcs represent transitions between environments. At any given time, one environment is selected as 'the current environment', defining a map of places with a certain topographical structure. One particular place within this map is selected as 'the current place'. These representations provide top-down biases on perceptual representations of the current environment. (And in situations where the observer is recollecting a distant environment, they fully recreate the recalled spatial structure of this environment.)

An important function of the environment/place system is to control *transitions* between environments. These transitions can involve sensorimotor operations, if the observer is actually moving through the physical world, or attending to different places within it. But alongside these sensorimotor operations, there are operations that update the current LTM environment—i.e. the currently active element within the graph representing all known environments—so that the LTM representation stays in synch with the physical reality. In addition there are operations that update the current LTM environment *without* any attendant sensorimotor operations, which allow the observer to disengage from his current physical environment and recreate distant environments from memory. In all of these respects, LTM representations play an important role in the environment/place system.

A significant feature of the LTM system is that it is *recurrent*: the operation that updates the current environment representation and establishes a spatially adjacent one involves following links from one environment representation to another environment representation *in the same medium*. This property of LTM environment representations relates closely to the deictic nature of representations in the environment/place medium, in ways which I will explain below.

2.2 Preliminaries: spatial representations in the 'reach' and 'grasp' visuomotor pathways

The account of spatial representations that I develop in this chapter draws on an account of spatial representations in the visuomotor pathways subserving reach-to-grasp actions that I have reviewed in detail elsewhere (Knott, 2012). In this section I will briefly describe the spatial representations computed in these two pathways, to provide background for the more general account developed in later sections.

Parietal cortex and premotor cortex commonly taken to compute a number of 'maps' of regions of space close to the observer (for reviews, see Colby and Goldberg, 1999; Andersen and Buneo, 2002; Batista *et al.*, 2007; Knott, 2012). While these maps are very hard to disentangle, it is useful to model them somewhat abstractly by assuming that different maps are centred on different effectors, or (more precisely) on different effector-based motor systems. In this simplification, a map for a given effector defines a region of space around or near this effector. Places in a given map are specified in a motor coordinate system, in which the coordinates are commands to the muscle groups within the relevant motor plant. So, if the motor system involves n independent joints, places in the map will be defined within a coordinate system with n dimensions.[1]

The best-studied motor maps are those related to the 'reach' motor system, that transports the hand to a given location in peripersonal space, and to the 'grasp' motor system, that orients and preshapes the hand for contact with objects and surfaces. I will review these in turn.

2.2.1 Motor maps in the reach pathway

The 'arm' motor system is most often studied in tasks involving reaching for target objects. In these tasks, the reach pathway in parietal and premotor cortex features several maps of locations around each hand, centred on the location of the hand (see e.g. Batista *et al.*, 2007; Brozzoli *et al.*, 2012). In a typical motor map, the position of an object directly above the right hand will be defined as the command which brings the hand into contact with this object. Importantly, the location of an object in the observer's peripersonal space in a motor map centred on a given hand will change *if the hand changes position*, even if the object itself does not move (see e.g. Graziano *et al.*, 1997).

[1] In fact, given that a motor command may have to bring a given joint into a given goal state by passing through a particular intermediate state (to generate an appropriate approach trajectory), there may be more than one point to define for each joint. There are various accounts of how these intermediate states are learned (see e.g. Oztop *et al.*, 2004; Lee-Hand and Knott, 2013); however, I will ignore the issue of via-points in the present discussion.

2.2.2 Motor maps in the grasp pathway

Parietal and premotor cortex also contain a variety of motor maps centred on the wrist, for orienting and preshaping the hand (see e.g. Murata *et al.*, 2000). These are again typically studied in the context of reaching-to-grasp. The grasp pathway running from parietal to premotor cortex computes the set of possible grasp affordances of an attended target object (see again Knott, 2012 for a detailed description). The wrist-centred motor maps computed in this pathway do not exactly represent the location of *points* in peripersonal space; they are best thought of as representing the location of *opposing pairs of surfaces* on an attended target object (see Iberall and Arbib, 1990). Iberall and Arbib define a pair of opposing surfaces within an object as an 'opposition space'. I will call it an *outer* opposition space, to keep room for a related term to be introduced later. The wrist-centred maps of space in the parietal and premotor cortex represent the set of opposition spaces within a selected target object, as a set of alternative goal configurations of the wrist, palm, fingers and thumb.[2] Importantly, these motor maps are all deictically referred to a single selected target object for which a reach action has already been computed in the 'reach' pathway. It is only after a single object is attentionally selected as a target in the reach pathway that visual attention is focussed on this object—and visual attention is necessary to generate the representations of object shape from which the grasp affordances are produced (see e.g. Deubel *et al.* 1998). However, selection of a given opposition space from the set of alternatives modulates the representation of the chosen reach action, so that the arm approaches the target from a suitable direction (see Fagg and Arbib, 1998).

2.3 The basic concept of a surface, and its representation in somatosensory and visual cortices

My model of the environment/place system draws on the framework introduced by Gibson (1950), emphasising the primacy of surfaces in an account of cognitive representations, and especially on surfaces defined within the modality of touch. In this section I will introduce a primitive notion of **surfaces**, defined initially in the motor system and then in the visual system. Surfaces will be the atomic units of representation within the environment/place system.

2.3.1 The stable support and contact signals

My main proposal is that the basic concept of surface is due to a primitive sensorimotor concept called the **stable contact signal** (and a special case of this signal called the **stable support signal**), which play a role analogous to the sensation of a stable grasp in my account of reaching-to-grasp. These signals are particular combinations of states in the sensorimotor system, which I assume are axiomatically rewarding in certain contexts:

[2] Or more precisely, as a set of alternative *trajectories* of these effectors, each *culminating* in a given goal state.

I propose that they automatically trigger various kinds of learning in infants during development, which bootstrap both motor and perceptual concepts of a **surface**. In this section I will describe the form of the stable contact and stable support signals, first in actions involving the hand/arm system, and then in the motor system controlling the observer's posture in his environment.

2.3.2 The stable contact and stable support signals in the hand/arm motor system

I will first introduce the stable contact and stable support signals informally, and then discuss how they are represented in the somatosensory cortex.

To define the sensation of stable contact in relation to the hand/arm effector, this effector should be thought of as a motor system, capable of various types of movement, and of receiving various types of perceptual input, including touch and proprioception. I propose that the stable contact signal is activated within this motor system when a particular combination of perceptual inputs and motor contingencies is detected. These are as follows. Firstly, there is a stable tactile signal registered in somatosensory cortex within one of the hand's component surfaces: for instance, the open palm, or the surface formed by the wrist and the fingertips. I will call this surface the **contact surface**, and define a vector normal to this surface in the direction of contact called the **support vector**. Secondly, the movements within the hand/arm motor system are constrained in various specific ways. To begin with, it is not possible to move the contact surface in the direction of the support vector: attempts to move in this direction are associated with an increased pressure on the contact surface. However, it is possible to generate motor commands that move the contact surface in directions orthogonal to the support vector. These movements should involve minimal changes to the pressure at the contact surface, and minimal changes to the configuration of the contact surface within the hand (i.e. minimal changes to hand position), together with the registering of a uniform 'slip' sensation across the whole contact surface. ('Slip' is delivered by a specialised class of mechanoreceptors in the skin; see e.g. Adams *et al.*, 2013 for a review.) All of these contingencies are combined into a single atomic neural signal within in the hand/arm motor system, that I will term 'stable support'.

The **stable support** signal is defined as a special case of the stable contact signal, in which there is additionally a change in the *weight* borne within the hand/arm motor system: specifically a reduction in its normal weight. This signal indicates not only that the hand is in stable contact with an external surface, but that this surface can potentially support the hand, or maybe even the whole body. Such surfaces tend to be horizontal, or close to horizontal. But contact with non-horizontal surfaces can also bear some of the effector's weight, especially if a horizontal component of force is being applied to the contact surface.

2.3.3 Representation of surfaces in the hand/arm somatosensory system

The stable support and stable contact signals for the hand/arm motor system are related to the modality of touch and proprioception: that means they are primarily represented in **somatosensory cortex**. In this section, I will introduce the idea that this area encodes representations of surfaces. I will discuss how somatosensory representations are learned in Section 2.3.5.1.1.

Somatosensory cortex is divided into 'primary' and 'secondary' somatosensory cortices (**SI** and**SII** respectively). The simplest somatosensory representations in SI encode a map of positions on the body, identifying the surfaces on the body where contact is made (see e.g. Penfield and Rasmussen, 1950). However, as with the parietal and premotor maps discussed in Section 2.2, there are more derived representations in somatosensory cortex in which the structure of this map is modulated by the current *pose* of the body (see e.g. Cohen *et al.*, 1994; Tillery *et al.*, 1996; Hamada and Suzuki, 2005; also Haggard *et al.*, 2006 for relevant behavioural data). These modulations allow SII to compute a representation of the surfaces *in the external world* that are contacting the body, as well as of the surfaces within the body that are being contacted (see Haggard, 2006 for useful discussion). This point is quite a subtle one, so it is useful to illustrate with a specific example.

I will assume a scenario in which there is a surface in the observer's peripersonal space whose location and orientation are fixed in relation to the observer's body, and which the observer 'explores' by making arm movements and detecting touches on his hands. (Importantly, I assume that the observer's movements don't cause the surface to move, even when his hands make contact with it.) Summing over the findings of pose-sensitivity noted above, we might consider a notional SII cell that responds when the observer's hand makes contact with a surface, *but only when his arm is in a particular position, and his hand is oriented in a particular way.*[3] For instance, such a cell might be active when there is contact on the observer's palm, but only when his arm is pointing straight out in front of him and his palm is oriented downwards. We can therefore speak of 'receptive fields' of SII cells *in peripersonal space*, as well as in body-centred space. For many SII cells, receptive fields are better described in relation to peripersonal space than to body parts, because they respond to contact on several *different* body parts. For instance, there are SII cells that respond to touches by several different fingers (see e.g. Fitzgerald *et al.*, 2006), or even to touches by both hands (see e.g. Iwamura *et al.*, 2001). In these cases, the fingers are often aligned within a single plane in the hand (see e.g. Fitzgerald *et al.*, 2006). A conclusion drawn by several researchers, and well summarised by Haggard (2006), is that SII cells compute representations of surfaces in the observer's peripersonal space. These representations can be thought of as representations of stable contact or stable support in the somatosensory system. (Note that the somatosensory system also computes the weight borne by the hand/arm, which is important for defining stable support; see e.g. Lederman

[3] It is more likely that this kind of response is generated by a whole *population* of SII cells. But I will refer to a single cell in the discussion, for simplicity's sake.

and Klatzky, 2011.)

2.3.3.1 Representation of curved surfaces in somatosensory cortex

Many surfaces in nature are curved rather than flat. There is good evidence that SII computes representations of surfaces with different degrees of curvature: for instance, Yau *et al.* (2013a) found that macaque SII neurons encode surfaces with a variety of different curvatures. These representations presumably have their origin in situations where a curved surface in the body (and in particular in the hand) makes stable contact with external surfaces, and in the motor routines that preserve this kind of stable contact when the hand is moved.

Curved surfaces in the hand can be created in a variety of ways. I will illustrate with the most obvious surface within the hand: the plane made up by the palm and the inner surfaces of the fingers and thumb. This surface can be shaped to create planes with a range of curvatures in two orthogonal axes in the palm: one perpendicular to the fingers and one parallel with them. There are limits to the curvatures that can be achieved, naturally, and to how these can be combined. For instance we can create a tight convex curvature in the axis perpendicular to the fingers (by curling the fingers), but only a very slight concave curvature in this plane. Or we can create a modest convex curvature in the axis parallel with the fingers—but not when there is also tight curvature in the other axis.

Motor states of the hand that create a *uniformly* curved surface within the hand form a special class of motor state in the haptic system, because there are natural circumstances in the world in which these states can be *maintained* during haptic exploration. If the surface being explored is also uniformly curved, around an axis aligned with the uniformly curving surface in the hand, then the curve in the hand can be maintained as the hand is moved if the hand travels along the surface in a direction perpendicular to the axis of curvature. Therefore there is a special class of stable contact signal activated when the hand is travelling along a surface with uniform curvature.

The fact that the hand can create uniformly curving surfaces in two dimensions, together with the idea that external surfaces are defined using the notion of stable contact defined above, means that the haptic system centred on the hand can readily describe a particular class of simple curved surfaces in the world, with uniform curvature in one or two dimensions. If curvature of the surface is zero in both dimensions, the surface is a flat plane. (We can think of a flat plane as a special case of a curved surface.) If curvature is convex and equal in both dimensions, the surface is the outside of a sphere. If if it is zero in one dimension and convex in the other, the surface is the outside of a cylinder. If curvature in both dimensions is convex, but the degrees of curvature are sufficiently different, the surface is the outside of a ring torus.[4] In Section **??** I will argue that surfaces of these types, with uniform curvature in one or two dimensions, provide a natural set of haptically-derived primitives for representing the 'parts' of more complex shapes.

[4] It is also possible to create surfaces with uniform curvature in the hand in non-orthogonal axes. These define curved surfaces with *torsion*, which can also be found quite commonly in nature, but for simplicity's sake I won't consider them here.

Other hand states, for instance a state with alternating raised and lowered fingers, do not permit this type of navigation: if each finger happens to be touching some portion of a complex surface, the only way to smoothly slide the contact surface of the hand over the object is to reshape the hand in very complex ways while it is moving, and this violates the definition of stable contact.

2.3.3.1.1 A coordinate system centred on the hand These two axes define a natural coordinate system for the hand. The axis parallel with the fingers provides a natural 'forward/back' axis, with 'forward' being the direction in which the fingers point. The axis perpendicular to the fingers provides a natural 'left/right' axis. I will assume a coordinate system centred on the right hand: in this coordinate system, 'left' is the direction which the chopping edge of the hand faces. I will assume these axes lie within the surface of the palm, so they curve when the palm curves (and extend beyond the hand with the same degree of curvature).[5]

It is also useful to define a third direction, 'up/down', that is perpendicular to the plane of the hand. This is not an axis as such, because its direction varies for different points in the hand, but it is useful for defining notions in haptic navigation. 'Down' is the direction of the 'support vector' in the terminology introduced in Section 2.3.2 (i.e. the direction in which the surface applies force to the support surface of the hand). 'Up' is the direction of movement which will break contact with the surface.

2.3.4 The stable support signal defined for the observer's whole body

Some surfaces are surfaces we can explore with effectors like hands; others are surfaces we can explore by locomotion. In this section I will briefly outline a concept of stable support which allows representations of the latter kind of surfaces to be learned. I will go into more detail about the motor systems involved in Section 5.1.2.

The stable support signal in the case of an observer's whole body is basically the sensation of being *balanced*. This has various components. One is vestibular: the observer should feel that his body is not moving, or at least not rotating about a horizontal axis.[6] Another is sensorimotor: there should be a motor system within the body, which I will call the **base motor system**, which is stably supporting the body's weight. Within this system there should be a support surface, somewhat analogous to the support surface in the hand/arm system, which should receive a constant tactile signal.

The support surface of the body is largely defined by gravity: it will typically be horizontal, or mostly horizontal. But there can also be non-horizontal support surfaces for

[5] Naturally these axes only correspond to the body's own foward/back and left/right axes in one particular hand/arm state. If there is any risk of ambiguity, I will use the terms 'forward$_h$', 'back$_h$, etc.'

[6] Visual texture flow can deliver similar information—although it is alright for the observer's *head* to be rotating horizontally, provided his body is not doing so.

the body: for instance vertical walls that can be leaned against. These too can contribute to an agent achieving the state of balance.[7]

Like the stable support signal in the hand/arm motor system, I assume that the 'balance' signal is intrinsically rewarding. It teaches infants a variety of ways to maintain balance in their environment (typically on a horizontal surface of some kind), and also ways of moving in the environment that maintain this balance. Equally importantly, it allows infants to learn perceptual representations of surfaces within their environment—both horizontal surfaces and vertical ones—as I will discuss in Section **??**.

I will discuss the base motor system in much more detail in Section 5.1.2: the main point of the current section is just to suggest that the concept of stable support can serve as the basis for representations of surfaces quite generally, not just representations of the surfaces of manipulable objects.

2.3.5 Learning perceptual representations of surfaces

When we look at our local environment, our spatial perception capabilities identify the surfaces that are present within it, in various different ways. In my model of spatial representations, the concept of a surface is originally defined in somatomotor terms. In this section I will describe how these basic somatomotor representations of surfaces can be mapped to visual representations, to enable us to represent not just the surfaces we are currently touching, but a much larger array of distant surfaces in our local perispace and environment. In Section 2.3.5.1 I will consider representations in the hand/arm motor system; in Section 2.3.5.2 I will consider representations in the navigation system.

The key idea in my learning model is that the stable support signal is intrinsically rewarding, at least in developing infants: in other words that it is hard-wired by evolution to generate a reward signal in the motor system. This means that infants are drawn to learn how to achieve stable support states, and consequently, to learn functions mapping perceptual representations of objects in their peripersonal space onto goal motor states associated with the stable support signal. These are at the origin of our perceptual representations of the surfaces of objects.

2.3.5.1 Learning surface representations in the hand/arm motor system

I will begin in Section 2.3.5.1.1 by how purely somatosensory representations of surfaces might be learned. In Section 2.3.5.1.3 I will consider how visual representations of surfaces can be learned using somatosensory representations as training signals.

[7] In fact, vertical surfaces are important for balance even if the observer is not in contact with them. Specifically, a vertical surface directly in front of the agent generates a visual representation that is used to control 'sway' (see e.g.): expansion and contraction of optic texture control sway forwards and backwards, and horizontal optic flow controls sway from side to side. (I think this may mean that when the agent switches his visual attention from one vertical surface to another, as often happens when he re-orients to a new boundary pair, there is a very small 'loss of balance' that may a kind of reconfiguration action comparable to the kind you get in the haptic system when you move a hand surface from one surface in an object to another. Not sure this is the right place to say this, though.)

2.3.5.1.1 Learning somatosensory representations of support surfaces for the hand/arm Recall from Section 2.3.2 that somatosensory cortex (and especially SII) can be thought of as computing touch-derived representations of the surfaces in the observer's peripersonal space. These representations are of 'whole surfaces', not just of the points in space with which the observer is *currently* in contact. As suggested in Section 2.3.2, the primary purpose of these surface representations is to support the movements that *explore* surfaces: that is, movements that maintain a constant stable support signal while traversing the surface. My proposal is that SII gradually learns representations of the common types of surface encountered by the observer in his environment, and that these help guide his exploration of new surfaces that he encounters.

How can these representations be learned? I will outline a simple learning routine which generates representations of commonly-encountered surfaces from tactile inputs in somatosensory cortex, enabling this kind of navigation. The architecture is illustrated in Figure 2.1.

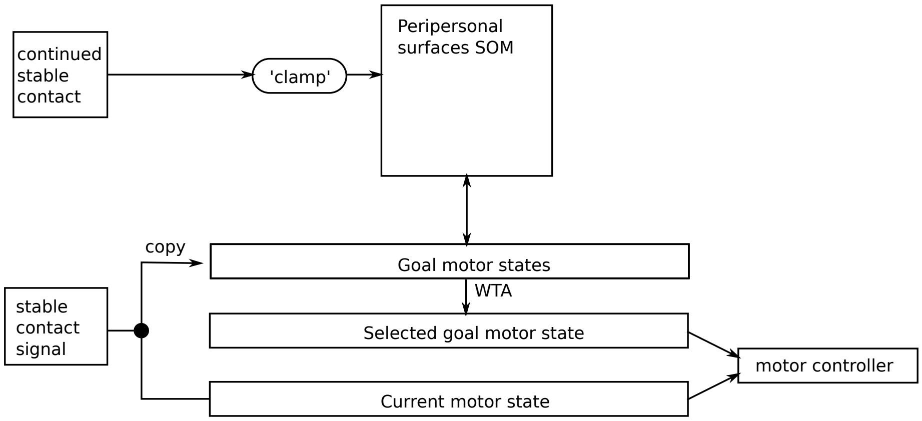


Figure 2.1: A schematic architecture for learning somatosensory representations of surfaces

In this figure the three lower media each encode all possible motor states of the hand/arm. The 'current motor state' layer holds a representation of the current static motor state of the hand/arm, delivered bottom up by proprioception (e.g. from stretch receptors in the joints of the hand, wrist and arm). The 'goal motor states' and 'selected goal motor state' layers are isomorphic to the current motor states layer, but are used top-down, to encode goal motor states to attain. The goal motor states layer represents a set of possible motor goals, from which a single goal is selected using a winner-take-all mechanism. The current motor state and the selected goal motor state provide input to a motor controller in premotor/motor cortex that moves the hand/arm towards this goal state. Finally, the 'peripersonal surfaces SOM' layer is a **self-organising network** (**SOM**) that learns to represent all surfaces encountered in the observer's peripersonal space during training, emphasising the types of surface that are encountered most often.

During training, the observer moves his hand/arm at random until a stable support signal happens to be activated. At this point, the current motor state is copied to the goal motor states layer and activated by itself, a unit is selected in the peripersonal surfaces SOM to represent the surface being touched, and the connection between the newly active units in the goal motor states layer and the SOM is strengthened. (The SOM unit is selected as a function of the unit in the goal motor states layer, plus a component of noise that initially dominates and is gradually annealed to zero during training). After making contact, the observer executes a *subsequent* movement. This subsequent movement is selected as a function of the currently active SOM unit, plus another noise component annealed to zero during training, so initially subsequent movements are generated at random. The subsequent movement results in a new proprioceptive hand/arm state. Crucially, if the subsequent movement results in *continued* contact with the surface (which will be registered as an unchanged stable contact signal, together with a 'slip' sensation), the SOM unit used to represent this state is 'clamped'—in other words, constrained to be the same SOM unit that was previously active. In this case, additional subsequent movements continue to be generated until contact is lost with the surface.

After the network has been trained on a range of different surfaces, and its noise components have been annealed to zero, SOM units come to represent different types of surface. Each SOM unit that was activated during training is potentially linked to a *set* of hand/arm states in which the hand is in contact with a particular surface. Surfaces in different parts of peripersonal space, and with different orientations, will be represented by different units in the SOM. Now, when the hand explores its perispace and contacts a surface, the SOM unit representing that surface will be activated. This unit will be linked to a *set* of units in the goal motor states layer, which each represent states in which the hand is in contact with the surface. To explore the surface, the observer can choose one of these as his new motor goal. If hand/arm states that are close to the current state are preferred, he will generate smooth actions sliding the hand along the surface.

My suggestion is that SOM units in this network play a role in behaviour somewhat analous to that played by SII cells encoding surfaces in somatosensory cortex. Particular SOM units function to organise points in motor space into planes with different locations and orientations. These provide the foundation for more complex representations of surfaces, both in somatosensory cortex and in visuomotor pathways.

2.3.5.1.2 Learning hand-centred navigation movements As well as learning to represent different types of surface, the observer's haptic experience with surfaces can teach him general principles about the movements of the arm (and other motor systems) that move the hand *within its own coordinate system* along a surface currently being contacted.

As mentioned in Section 2.3.3.1, the hand's surface can readily be curled along two roughly orthogonal axes: a 'forward/back' axis aligned with the fingers, and a 'left/right' axis perpendicular to this. These axes can be extended beyond the plane of the hand (maintaining its curvature) so they define a contact surface with constant curvature within which the hand can *travel*. In other words, surfaces in the world that have constant curvature can have their own intrinsic geometry described in the coordinate system of the hand.

In order to make his hand 'navigate' within a coordinate system centred on a surface with constant curvature, he must learn what motor movements will cause the hand to travel along such surfaces in the 'forward', 'back', 'left' and 'right' directions. The learning process will require a lot of experience with objects with different curvatures in different peripersonal locations. The learned movements will move the hand within peripersonal space, and for curved surfaces, rotate the hand appropriately while it is moving. I will not consider in any detail how this learning takes place. I will simply assume that it can happen—in other words, that the observer can learn to make his hand move in arbitrary directions around any surface with constant curvature in two orthogonal dimensions, in a coordinate system centred on these two dimensions.

2.3.5.1.3 Learning visual representations of support surfaces for the hand/arm

I will now consider how the observer can learn to generate visual representations of surfaces that can be reached by the hand/arm motor system. In this case, the main relevant pathway is the one running from early visual areas through posterior parietal cortex to premotor cortex, that generates motor affordances of visually perceived stimuli, as discussed in Section 2.2. There is good evidence from many sources that representations of surfaces and shapes in this pathway are activated by both vision and touch (see e.g. Lacey *et al.*, 2009 for a review).

There are many types of motor affordance. The account of reaching-to-grasp in Section 2.2 focussed on the generation of *grasp action* affordances. Now I will consider the affordances of another class of actions, which achieve stable support relationships with surfaces in the observer's perispace. These are not quite the same: the actions that achieve a stable grasp of a book are not the same as those that represent the book as a surface. But the principle by which support affordances are learned is very similar.

The framework for the learning model is that posterior parietal cortex implements a function mapping the stimulus at the attended region of the visual field onto a set of **goal motor states** of the hand/arm motor system. Each of these goal motor states is associated with a reward signal. In the case of reach-to-grasp actions, the reward signal is the sensation of a stable grasp. In the case of actions that bring the hand into a stable support relation with a nearby surface, the reward signal is the stable support signal. What we have to explain is how the function mapping visual representations to support affordances is learned.

The basic training routine is similar to that outlined in Section 2.3.5.1.1: the observer executes hand/arm movements towards randomly selected points in peripersonal space. But this time he also attends visually to the selected region of space in each case. (To explain how this is possible, we assume a model in which the 'reach' component of the movement is generated from a retinal representation of the most salient item in the visual field; see Lee-Hand and Knott, 2013.) From time to time, as before, the random movement results in activation of a stable support signal. This reward signal triggers an operant learning operation. First, the *current* motor state of the hand/arm is copied into another medium holding the observer's *goal* motor state. Next, the visual stimulus at the attended location is trained to map to this goal motor state. The relevant properties of the visual stimulus will be those that identify its distance, and its local planar orientation. (These visual properties are known to be computed in the area of parietal cortex controlling hand/wrist movements, anterior intraparietal cortex; see e.g. Sakata *et al.*, 1998; Murata *et al.*, 2000; stereopsis is known to play an important role in their computation.) With enough learning of this kind, the trained function will map the distance and local orientation of the surface at each point in the observer's perispace projecting to the retina onto a goal motor state in which a stable support relationship with the surface at this point is achieved. (I assume a general motor controller that can bring the hand/arm into arbitrary goal motor states, as in many computational models of motor control; see e.g. Jordan and Wolpert, 2000).

We can also envisage a function that maps more holistic visual representations of surfaces directly onto somatosensory representations of whole surfaces in SII (i.e onto representations analogous to the SOM units in Section 2.3.5.1.1). This may be particularly helpful in cases where the surfaces with particular shapes have distinctive visual forms—for instance in the case of surfaces with different curvature. Recall from Section 2.3.3.1 that Yau *et al.* (2013a) found macaque SII neurons that encode the curvature of a touched surface. Interestingly, the same group found neurons in the visual area V4 of macaque that encode perceived surfaces with specific curvatures (Yau *et al*, 2013b). (Similar representations of three-dimensional visual curvature have been found later on in the macaque dorsal visuomotor pathway, for instance in anterior intraparietal cortex; see e.g. Theys *et al.*, 2012). The curvature of a surface in a local region relates closely to its shape, so it makes sense to envisage that a holistic representation of shape in a given region is mapped directly to a somatosensory representation of a curved surface in that region. A final interesting case is when a flat surface is bounded by an edge: in this case, information about the 3D orientation of the surface can be read from the 2D boundary projected onto the retina, and there is a great deal of evidence that neurons in all visual pathways are sensitive to 2D orientation (see e.g. Yamane *et al.*, 2008 for references). Obviously, visual analysis of object shape does not just relate to the curvature and orientation of surfaces; I will consider visual representations of object shape more generally in Section 2.10. For the moment, my main point is that we can envisage simple visual representations of surfaces being learned by functions mapping both 2D and 3D visual representations onto representations of surfaces in somatosensory cortex, and that these functions are likely to learn a mixture of holistic and point-by-point mappings.

After learning is complete, the functions sketched above can be thought of as generating effective visual representations of the local orientation and curvature of support surfaces in the agent's perispace. They are effective because they connect to the motor system: they support the execution of actions that make contact with support surfaces, and then 'navigate' within them. I use the term 'navigate' deliberately, to highlight similarities between the representations of surfaces that can be explored with effectors like the hands and surfaces that can be explored by locomotion actions. To illustrate: an idea which will become important later on is that a surface which generates support affordances for the observer is also a place that the observer could potentially *be in*. The observer is certainly not in this environment yet. And even if the observer places an effector (like a hand or foot) on the surface, to achieve stable support on it, he is not in that environment (though his hand or foot will be). However, he may now be in a position to *transfer his weight* to the newly-placed effector, in which case he will establish the surface as his new environment. This idea will be introduced we we discuss operations that transition between environments in Section 2.18.1.

2.3.5.2 Learning surface representations in a locomotion motor system

At a first approximation, the surfaces that are relevant to the locomotion system are the horizontal surfaces that can be walked over (such as the ground or the floor), and the vertical surfaces that are barriers to locomotion (such as walls or cliffs). Rich perceptual representations of these surfaces can also be learned. These will be the focus of Section 2.4. In the current section I will consider some of the simpler perceptual representations which provide the atomic components of the richer representations discussed later, and which relate directly to a stable support signal.

2.3.5.2.1 Visual representations of navigable horizontal surfaces An important visual representation supporting navigation is a representation of the horizontal surface the agent is on. This representation is produced partly from stereopsis, which can compute distance and local orientation for surfaces reasonably close to the observer, as discussed in Section 2.3.5.1, and partly from a visual cue called texture gradient for horizontal surfaces further away.

Assume that the observer is at a given location in an environment, and has some rudimentary ability both to maintain balance and to move to adjacent locations. For instance, assume the observer is an infant who can maintain balance on all fours, can turn to orient to different points in the environment, and can crawl forward. We can envisage a way for the observer to learn a simple visual representation of the navigable support surfaces around him which has a similar structure to the method described for hand actions in Section 2.3.5.1. The observer first generates a representation of the surfaces subtending the full visual field, and stores this. Then he executes a random locomotion movement, involving a turn and a crawl forward. This movement is extended in time: we can think of it as having several steps. At each step, the observer learns a function mapping the stored representation of surfaces onto the sequence of movements made so far. Learning stops in various circumstances. One is when locomotion becomes impossible, because of a barrier such as a wall, or an obstacle. Another is when balance is lost, for instance in the case where the observer encounters a cluttered or slippery surface (or worse, navigates over a drop, and falls). Another is when the stored representation fades, and learning times out. After this form of learning, the agent will learn to generate a representation of the navigable horizontal surface within his visual field, that supports the planning of a range of locomotion actions. The boundaries of the navigable surface will also be identified, in a negative sense: the learned function will simply not produce actions that go beyond the navigable surface as candidates (whether this is because they encounter a boundary or obstacle, or result in loss of balance). Of course, this representation of the observer's navigable surface is defined in retinotopic terms, and is thus highly inefficient. However, it can provide useful input to functions that re-represent these boundaries more abstractly. These are the functions to be discussed in the following sections.

Note, incidentally, that the navigable support surface defined in the above scheme can also be defined as a support surface for the agent's effectors—at least locally, near the agent. For instance, a crawling baby can explore the ground as a surface that affords stable support for his hand, if his hand is represented as an effector, rather than as part of the balance system. (This of course requires the baby to balance on knees and a single hand, while the other hand is exploring the floor.) The important thing is that the surface defined in hand-centred terms *coincides* with the surface defined in relation to navigation actions. This notion of surfaces defined in multiple converging frames of reference will recur throughout the environment/place system.

2.3.5.2.2 Visual representations of vertical boundary surfaces Before we consider these more abstract functions, it is useful to note that the observer can also generate 'positive' retinotopic representations of the edges of the navigable environment, if these are surfaces in their own right. Again, these can be learned within a system gated by activation of a stable support signal in the motor system concerned with balance. Recall from Section 2.3.4 that a vertical surface such as a wall can also contribute to an agent's balance: if a wall is nearby, the observer can learn new ways of balancing, involving *leaning*: one can lean sideways, backwards or forwards, making contact against the wall with the body, or with arms, or outstretched hands. This kind of contact generates a stable support signal that is quite distinct from the one that comes from unsupported balance. We can envisage a separate function that maps a representation of the surfaces in the visual field onto actions that generate this special kind of contact. These actions will also typically be locomotion actions—their special property will be that they *culminate* in a certain kind of balance state, involving contact with a vertical surface. Again, it is clear that these retinal representations will not generalise well—but again, they can serve as input to functions that learn more stable representations.

2.4 Representations of the current locomotion environment and places within it

In this section and Sections 2.5 and 2.6 I will introduce the two basic components of the environment/place system as it is used to represent the observer's current locomotion environment: a circuit for representing the environment holistically, and a circuit for representing places *within* the current locomotion environment. A key feature of the architecture I propose is that representations of places within the current environment are deictically referred to the currently active holistic representation of the environment. A similar scheme will be proposed for spatial representations of objects in Section 2.9.

The spatial representations described in these sections involve a mixture of directly perceptual representations and representations in a visuospatial working memory medium that integrates information received over a short interval of time. I will also make occasional reference to long-term memory representations, but these will be considered in their own right (see especially Sections 2.7 and 2.11).

2.4.1 Environment representations: the parahippocampal place area

A key structure involved in our representation of environments is the **parahippocampal place area** (or **PPA**). This area plays a role in representing the agent's current environment. In this section, I will summarise what is known about the representations it computes.

The key finding is that the PPA responds to 'environments', not objects. In the experiments which show this effect, beginning with Epstein and Kanwisher (1998), environments are defined as having an 'extended spatial layout', of the kind that an agent can be 'in', and objects are defined as having a 'compact spatial layout', and as being entities which an agent can act *on*. Examples of environments are landscapes, empty rooms, or urban scenes containing buildings and streets; examples of objects are faces, and household/office implements. fMRI experiments show that the PPA responds to environments but not to objects; importantly, it responds just as strongly to scenes containing no objects (for instance an empty room) as to scenes containing objects. Interestingly, it responds to scene-like stimuli even if they have a spatial scale which would not permit an agent to navigate them. For instance, a scene consisting of an arrangement of objects on a desktop evokes a good response in the PPA (Epstein *et al.*, 2003), as does a scene consisting of a lego model of walls and corridoors (Epstein *et al.*, 1999).

The PPA's representation of an environment appears primarily to encode the spatial layout of the surfaces which objects can rest on within it, and of the boundaries which delimit it (see also Park *et al.*, 2011). Epstein refers to these surfaces as 'background elements' of the scene. Note that these are exactly the kinds of elements that are identified by the perceptual representations of surfaces discussed in Section 2.3.5.2. If an observer is shown one scene twice in succession, the PPA responds less the second time, while if two different scenes with different spatial layouts are shown, the PPA's response to the second scene remains strong (Epstein *et al.*, 1999). The same paradigm also shows that the PPA responds differently to different *views* of a single scene (Epstein *et al.*, 2003). However, this viewpoint-specificity reduces as an observer gains experience about how a scene is spatially situated within a wider environment, at least in 'good navigators' (Epstein *et al.*, 2005.)

The PPA responds to both familiar and unfamiliar environments. However, it responds more strongly to familiar ones, which suggests that it has some role in the recognition of specific environments (Epstein *et al.*, 2007)—in other words, it is involved in storing or accessing representations of environments in long-term memory. However, Epstein *et al.*

(2007) also found that another parahippocampal area called the **retrosplenial cortex** was *more* sensitive to the distinction between familiar and unfamiliar environments. They propose that both areas have a role in recognising specific environments. Their suggestion is that environments are identified in part by their spatial layout, but also by their spatial relationships to other known environments, with this latter component being represented in retrosplenial cortex. I will return to this idea in Section **??**.

In line with the results just summarised, damage to the PPA or to the retrosplenial cortex results in impaired navigation abilities (Mendez and Cherrier, 2003). However, the PPA at least does not appear to be involved in *online control* of navigation, because it does respond any more to 'subjective-motion' environments than to static ones. (I will look at the perceptual routines involved in online control of motion in Section **??**.)

Importantly for the model I will outline, the PPA and the retrosplenial cortex represent an *attended visual region*. The experiments described above present an environment to a viewer on a computer monitor; the environment to which the PPA responds only occupies a small portion of the observer's visual field. If the PPA responded to the observer's *actual* environment, in which the computer monitor is simply an object, we would not expect to see any changes in its activity as the monitor displays pictures of different scenes. The fact that we do see such changes means that input to the PPA can be gated by visual attention, in the same way that input to the object classifier can be.

2.4.2 A map of places in the hippocampal region

2.4.2.1 Place cells and view cells

Several very well-studied populations of cells in the hippocampus and surrounding regions appear to encode maps of places within the agent's local environment. One of these is the ensemble of hippocampal **place cells** (O'Keefe and Nadel, 1978; see Jeffery, 2011 for a recent review). Place cells represent the agent's own location in an allocentric reference frame (i.e. one that is centred on the environment, rather than on the agent himself). When the agent moves around his environment, a given place cell is mostly quiet, but it fires when the agent is in a particular region, called that cell's **place field**, regardless of the direction from which this region is entered.

Place cells are found in all mammals, although they are most frequently studied in rats. Another population of place-encoding cells in the hippocampus are termed **view cells** (Rolls *et al.*, 1997; Rolls, 1999). These represent the location of an external object which is currently being attended, rather than the location of the agent himself, again given in coordinates centred on the local environment. A given view cell will fire whenever the agent atttends to an object in a particular location in the environment (again termed this cell's place field), no matter where the agent is in relation to this location. View cells are only found in primates.

It is interesting that the 'local environment' mapped by place and view cells is defined similarly to the environment to which the parahippocampal place area responds. In each case, the local environment is typically an enclosed area, whose shape is defined by a collection of *boundaries*. Experiments investigating spatial representations in the hippocampus are frequently conducted on rats, in which case the boundaries tend to be the walls of a maze. But boundaries can also be created by a cliff, or a body of water, or an area of unnavigable terrain.

2.4.2.2 Grid cells

A final population of place-encoding cells in the hippocampal system are **grid cells**, which are in the entorhinal cortex rather than the hippocampus proper (Fyhn *et al.*, 2004). While the place field of a place or view cell is typically a single contiguous region, a typical grid cell has multiple place fields, arranged in a hexagonal pattern covering the whole environment. Grid cells are anchored to external landmarks, but they do not encode the spatial structure of the environment in any detail: their main role is probably to support 'dead reckoning'—that is, updates to representations of place in the hippocampus as the agent moves through the environment.

2.4.2.3 Representation of trajectories the place cell system

Recall from Section 2.3.5.2 that locomotion actions can be thought of as sequentially structured: for instance as sequences of steps. There is mounting evidence that ensembles of hippocampal place cells can encode *sequences* of locations, as well as individual locations. Frank *et al.* (2000) found that the firing of a hippocampal place cell sensitive to a given location was modulated by the path the animal had taken to arrive at that location, and also by the path the animal subsequently took. Ferbinteanu and Shapiro (2003) found similar prospective and retrospective sensitivity to the animal's trajectory in hippocampal cells. In addition, they found that prospective encoding was diminished in trials where the animal made a navigational error, suggesting that this type of encoding is involved in the animal's goal-oriented navigational behaviour. More recently, Ji and Wilson (2008) studied rats in the process of switching from a well-learned trajectory to a new, partially overlapping trajectory. They found that learning the new trajectory caused the activity of cells associated with the overlapping region to become increasingly sensitive to past locations on the new trajectory. This suggests a mechanism for learning a new trajectory, involving the creation of new dependencies between cells encoding successive positions on the trajectory. Ji and Wilson also found that the changes associated with new trajectory learning preceded a reliable behavioural switch to the new trajectory—again, this suggests that the changes have a causal role in influencing the rat's behaviour. In summary, there is good evidence that trajectories are a natural unit of information encoding in the hippocampus.

2.4.2.4 Head direction cells

As already mentioned, hippocampal representations of places in the agent's current environment do not encode information about the agent's orientation. This information is represented separately, in a population of **head direction cells**, which encode the direction an agent's head is pointing in, again in a coordinate system centred on the local environment. Head direction cells are found in the postsubiculum (see Taube *et al.*, 1990) and in several other areas, including retrosplenial cortex (see Taube, 2007). They represent the agent's head direction, but not his location; thus place cells and head direction cells provide two quite separate components of the agent's current allocentric spatial state.

2.5 Boundary structures for encoding the relationship between environments and places

In this section I will outline a suggestion about how hippocampal maps of places relate to environment representations in the PPA. The key linking role is played by representations of boundaries.

2.5.1 The structure of the map of places: topographical or topological?

There is a good consensus that the system of hippocampal place cells introduced in Section 2.4.2.1 represents aspects of the spatial structure of the agent's local environment, in a way that supports navigation within this environment (see e.g. Nadel, 1991; Maguire *et al.*, 1998). However, how it does so is still a matter of debate.

The two main proposals draw inspiration from the same experimental findings: in particular the finding that place fields of place cells deform as a function of the geometry of the agent's current environment. In an early demonstration of this principle, O'Keefe and Burgess (1996) placed rats in a corridor environment whose walls could be made narrower or wider: they found that place fields of hippocampal place cells 'squash' and 'stretch' in a direction perpendicular to the walls. This finding has led to two quite different proposals about how hippocampal place cells encode the spatial structure of the environment.

In one model, place cells represent the agent's location as a function of his distance from and orientation towards selected boundaries of his currrent environment, so-called 'boundary vectors' (Barry *et al.*, 2006). In this proposal, place cells represent geometrical, or topographical, information, varying roughly monotonically as the agent moves in relation to environmental boundaries. However, there is also long-standing evidence that place fields can change in more radical ways in response to changes in the geometry of the environment: for instance, place cells whose fields overlap in a square environment might not have overlapping fields in a circular environment (see e.g. Muller and Kubie, 1987).

The other proposal is that hippocampal place cells represent the local environment using a topological scheme rather than a geometrical one (see e.g. Dabaghian *et al.*, 2014; Chen *et al.*, 2014). In this model, the environment is represented as a discrete set of places, connected by arcs encoding relationships of adjacency. This model originates in the observation that the neural media downstream of the hippocampus have no direct access to the place fields identified by neuroscientists, but they do have direct access to adjacency relationships between place cells: when the place fields of two place cells overlap, or are adjacent, they will fire simultaneously, or in succession. As discussed in Section 2.4.2.3, sequences of place cells are a natural unit of representation within the hippocampus: this finding supports a model in which sequences of hippocampal place cells represent topological structures within the current environment.

2.5.2 The mapping from retinotopic to allocentric representations of environments

Whether the spatial structure of the environment is encoded topographically or topologically, or using some combination of these methods, we have to envisage a function that *imposes* a specific encoding of the spatial structure of the current environment on the place cell system, which is different from one environment to another. Representations within the place cell system are classically deictic: they cannot be interpreted until we know what spatial structure that has been imposed on them. A key question is: how is a spatial structure imposed on the system of place cells? We must envisage another, more holistic, representation of 'the shape of the current environment', reflecting either the agent's current perceptual representation of the environment, or a representation of the environment retrieved from long-term memory, or both. Then we can envisage a function mapping this holistic environment representation onto a hippocampal state imposing a particular spatial structure on the system of place cells.

A likely candidate for the holistic environment representation providing input to this function is the parahippocampal place area (see e.g. Doeller and Kaplan, 2011). As discussed in Section 2.4.1, the PPA represents the spatial layout of an observed environment in an observer-centred coordinate system, largely by representing the configuration of its boundaries. But the PPA is also involved in *recognising* environments the observer is familiar with. (Recall from Section 2.4.1 that the PPA recognises environments by their spatial layout.) The PPA must therefore store a representation of the layout of each environment the agent is familiar with, as well as a representation of the currently perceived environment which can be matched to these. Since the observer can recognise a known environment from different perspectives, the process of recognising the current environment must involve generating a representation of its layout that abstracts away from the observer's current position and orientation in it. As already mentioned, the PPA's representations of known environments are more viewpoint-independent than its representations of novel environments (Epstein *et al.*, 2005): this is consistent with the idea that it has a role in converting an observer-centred representation of the perceived environment onto a viewpoint-independent one. Epstein *et al.* (2003) suggest that the PPA's egocentric representation of the spatial layout of the current environment has a central role both in defining the spatial structure currently in force in the place cells system and in generating environment-centred representations of the observer's current location and head direction within this system.

This is a useful proposal, and I will adopt it in the model I develop. But it is important to consider *how* the PPA's environment representations impose a spatial structure on the place cells system. I will make some suggestions below.

2.5.2.1 Retinotopic and allocentric boundary structures

Firstly, I propose that the PPA maps an egocentric representation of the boundaries in the environment onto an allocentric representation of these same boundaries. This idea is already implicit in the proposal of Epstein *et al.* (2003), but I will be somewhat more concrete. An allocentric representation of the boundaries of an environment is given in a coordinate system centred on that environment: for instance, centred on its major axis. It is not a representation of 'places in' an environment: it is just a representation of boundaries and the relationships between these. I suggest that each environment stored in long-term memory (LTM) is associated with an **allocentric boundary structure**. When the retinotopic boundary representation derived from perception is converted to an allocentric boundary structure, it can be directly compared with these stored allocentric boundary structures, and—at a first approximation—if there is a match, the boundary is recognised.[8] In turn, the allocentric boundary structure generated from perception, perhaps coupled with an allocentric boundary structure retrieved from LTM, imposes a spatial structure on the place cells system. Once this spatial structure is in place, the retinotopic boundary representation can be used again, to identify the observer's allocentric place and orientation within the newly-defined spatial structure of the current environment.

2.5.2.2 The form of boundary structure representations

What form do the 'boundary structures' representing environments in PPA have? There are indications from several sources that they do not have to represent all the boundaries in an environment at once. I will briefly summarise these.

Firstly, recall the boundary vector model of Barry *et al.* (2006). In this model, the spatial structure imposed on the place cell system is a function of the agent's position in relation to *selected* boundaries in the current environment. The place field of a place cell is most strongly determined by the boundary or boundaries which are close to it (though it is also sensitive to the distance to an 'opposing' boundary on the other side of the environment; see e.g. Sharp, 1997). Cells representing boundary vectors have been found in the subiculum of rats; whether they are found in the human PPA is not known, but the rat homologue of PPA, postrhinal cortex, is certainly one of the sources of input to the subiculum (Doeller and Kaplan, 2011).

Secondly, there is evidence that the system of hippocampal place cells represents different aspects of a complex environment in separate maps, and can switch very rapidly between alternative maps (see Derdikman and Moser, 2010 for a summary). As they note, hippocampal place cells have high variance in their firing rates—more than would be expected from a Poisson-like distribution. One possible explanation is that the hippocampus

[8] As mentioned in Section 2.4.1, environments are also defined in LTM through their relationships to other environments. This is another factor relevant to environment recognition, which will be discussed in Section 2.7.1.3.

is rapidly switching between several alternative maps, in which any given cell may play very different roles. There is some good evidence that such switching does indeed occur, and that it has its origin in the agent's attention to different aspects of the local environment (Fenton *et al.*, 2010).

Finally, if we think about the kind of semantic objects that can function as environments in linguistic representations, there is also good evidence that environments can be represented partially as well as completely. The prototypical environment is a three-dimensional space, like a room: this is the kind of environment that subjects are exposed to in Epstein *et al.*'s experiments. Looking ahead to Section **??**, I will propose that the spatial PP *in the room* denotes a process in the environment/place system in which the room is represented as an environment. But in language, smaller semantic objects can also be environments. As well as *in the room*, we can use expressions like *in the corner [of the room]* or *at the side [of the room]*. Corners and sides are also represented by boundaries, but they are representations of *selected* boundaries in an environment, rather than of all boundaries.

2.5.3 A model of allocentric boundary structures

In this section I will sketch a model of the allocentric boundary structures computed in PPA.

I will define an allocentric boundary structure as a set of pairs of roughly opposing boundaries. Each pair of opposing boundaries defines a spatial structure within the system of place cells: either a metric, topographical structure or a discrete topological one, or a mixture of these types (I will stay agnostic on this issue). I will assume that the representation of a pair of opposing boundaries also incorporates a representation of the vectors that link the ends of these boundaries, so that different degrees of overlap between the opposing boundaries can be represented, as well as boundaries of different lengths. Sometimes these vectors will correspond to real boundaries in the environment, for instance in the case where the opposing boundaries are connected by a wall; in this case, the connecting vector connects the boundaries at actual corners in the environment. In other cases, connecting vectors just define notional boundaries between sub-spaces in an environment. I will also assume that the opposing boundaries can be curved in various simple ways, analogously to the types of constant curvature discussed in Section 2.3.3.1, to provide components for representing environments with curved shapes. But in this section I will focus on simple environments with straight boundaries. I provide a more detailed discussion of the shapes of curved environments when considering the shapes of manipulable objects, see especially Section 2.9.5.7. A single boundary pair plus its connecting vectors defines a space that I will call an **opposition space**.

A **complete** allocentric boundary structure represents all the boundary pairs in an enclosed environment. But at any given time, I assume that a single boundary pair within this structure is **active**, and defines the current spatial structure within the system of place cells. In perceptual terms, the currently active boundary pair corresponds to the boundaries that are perceptually attended to at the current moment. Environments with very simple shapes, for instance simple quadrilateral environments, can be described by a single boundary pair. Environments with more complex shapes need to be represented by several boundary pairs, of which only one is active at a time.

An example of a (marginally) complex environment is shown in Figure 2.2: an L-shaped room. I suggest that the allocentric boundary structure representing this environment in

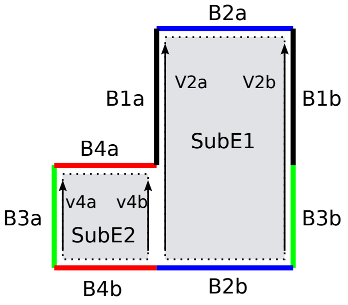


Figure 2.2: Proposed representation of an L-shaped room

LTM comprises four pairs of opposing boundaries, as shown in the figure, and several pairs of neighbouring boundaries. My proposal is that an observer cannot activate all these at once, but he can activate different pairs at different times. For instance, to represent sub-environment $SubE1$, he can activate the boundary pair $B2a$ and $B2b$, linked by the vectors $V2a$ and $V2b$; to represent sub-environment $SubE2$, he can activate the boundary pair $B4a$ and $B4b$, linked by vectors $V4a$ and $V4b$. (I assume that there are perceptual correlates of these operations focussing on particular boundary pairs, that focus attention on particular subspaces within the room.) There are also rectangular spaces defined by opposing boundaries $B1a/B1b$ and boundaries $B3a/B3b$.

2.5.4 Transitions between boundary pairs in an environment

In the case where an environment contains several boundary pairs, it is important that the observer can transition smoothly between different pairs. In the model I propose, the observer's representation of the shape of a complex environment is a dynamic entity, represented in a recurrent network. At any given time, there is a single active current boundary pair; the observer's knowledge of how the subspace represented by this boundary pair relates to the complete shape of the environment is reprsented by transition operations that activate different boundary pairs.

Transition operations exploit the fact that different boundary pairs overlap. For instance, in Figure 2.2, one of the vectors linking $B1a$ and $B1b$ is also one of the boundaries in the boundary pair $B2a$-$B2b$. A transition operation specifies the transformation that maps one boundary pair onto an overlapping one. This transformation may involve a rotation, or a figural change, represented in the relationships between the opposing boundaries and the vectors that connect them; it may also involve a change in absolute scale. For instance, when transitioning from the boundary pair $B1a$-$B1b$ to the pair $B2a$-$B2b$, there is a rotation, given by the angle between $B1a$ and $B2a$, and a change in the figure represented (the joining vectors and boundaries are of similar sizes for $B1a$-$B1b$, but the vectors are much longer than the boundaries for $B2a$-$B2b$).

In the remainder of this section, I will sketch a circuit that holds a representation of the shape of the observer's current local environment. An important preliminary point to note about this model is that it represents operations that transition between opposition spaces as operations within the *motor system*, that achieve high-level *motor goals*—specifically, navigational goals. In this conception, the observer's representation of his current environment is something that he can actively manipulate in service of his goals, rather than just a passive structure. There are three two levels of motor goals. A **goal place** is a place within the current opposition space that the observer wants to reach, for some reason. A **goal adjacent opposition space** is an overlapping opposition space that the observer wants to establish. A **goal local opposition space** is an arbitrary opposition space within the environment that the observer wants to establish. I will introduce the circuits that achieve these three types of goal one by one.

2.5.4.1 A circuit representing navigation actions within the current opposition space

A notional circuit for generating navigation actions within the current opposition space is illustrated in Figure 2.3.

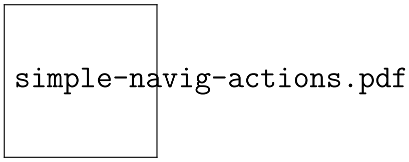


Figure 2.3: Sketched circuit for the navigation system

2.5.4.2 A circuit representing transitions between overlapping opposition spaces

The circuit is illustrated in Figure 2.4.

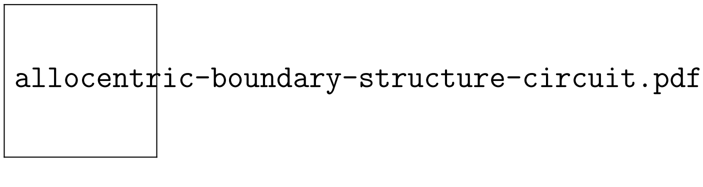


Figure 2.4: Sketched circuit implementing an allocentric boundary structure

At any given time within this circuit there is a single active **current opposition space**, which imposes a spatial structure on two isomorphic maps of 'places', representing the observer's **current place**, and the observer's **goal place**. The current place and goal place representations provide input to a navigation function that will be described in much more detail in Section **??**; for present purposes, the goal place is relevant because navigating from the current opposition space to an adjacent one involves moving to a particular subpart of the current space (the part that overlaps with the adjacent space). There is also a representation of the observer's **current orientation** within the current opposition space, and a representation of his **goal orientation**. Finally there is a single active **scaling factor**, that indicates the size of the current opposition space. This indicates the absolute length of a reference component of the current opposition space—say Ba—and is given in units defined within the motor system for navigation—say number of strides.

and a set of **candidate adjacent opposition spaces**, from which a **selected adjacent opposition space** can be picked.

The current opposition space imposes a spatial structure on the set of places. To get to a given

2.5.4.3 The role of goal opposition spaces in attentional actions

If we allow environments to contain multiple opposition spaces, then the mechanism that allows the observer to *attend* to other objects in his current environment becomes a little more complex. An object in the observer's current environment may not be in his current opposition space: in order to attend to such an object, the observer must potentially establish an opposition space other than the one he is currently in.

I will define an **orienting action** as a direction of attention of the observer that perceptually establishes an environment other than his own. I envisage that the process of attending to an environment is distinct from that of attending to an object, although the two processes are closely related. I will talk more about their relationship in Section **??**.

I suggest that attending to an environment is quite similar to actually entering that environment. Firstly, the observer must navigate to a position from which an operation transitioning to this environment is possible.

In the model I propose, the same LTM data structures

This notion requires an extension to the way environments are represented. In Section **??** I introduced a distinction between the observer and the currently-attended 'subject', who can either be the observer or an external individual. If the The observer must represent his own environment, but also the environment of the currently established subject

new data structure: as well as distinguishing between the current location and orientation of the observer and the location

This type of action relies on the notion of the current subject, which can be different from the observer.

In the simplest case, the observer just has to execute an attentional action. But in some cases, the observer has to move as well, to get into a place from which he can perform this attentional operation.

2.5.5 A circuit representing a complete environment

In order to represent a complete environment, it is necessary to represent relationships between arbitrary opposition spaces within it as well as relationships between adjacent spaces. For instance, an agent currently establishing $SubE1$ should know how to establish $SubE2$; this involves transitioning through the opposition space defined by boundaries $B3a$ and $B3b$. I suggest the observer stores transitions between arbitrary pairs of opposition spaces within his environment. In a small environment, transitions between all pairs of spaces are likely to be stored; in a larger ones, there may be some gaps in coverage, if the observer habitually travels through the environment in certain canonical ways. I propose that transitions between arbitary pairs of environments are represented as stored *sequences* of transitions between overlapping opposition spaces. Finally, I suggest it is important that an environment representation directly specifies all of its component opposition spaces, so that there is direct access to all of them.

There is also a set of **candidate local opposition spaces**, representing the complete set of opposition spaces within the environment, from which a **selected local opposition space** can again be selected.

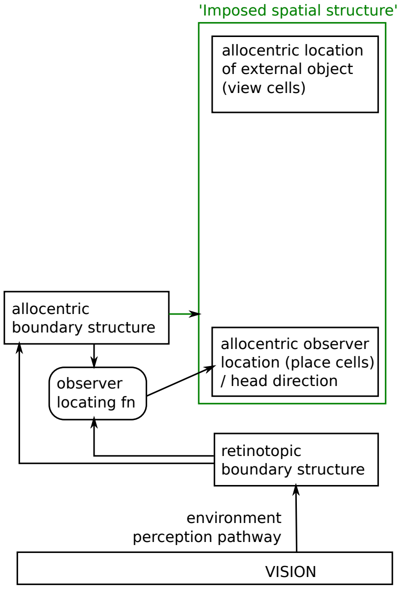
2.5.6 Boundary pair representations in a body-centred coordinate system

In Section 2.5.3 I suggested that pairs of adjacent boundaries in a boundary structure are defined by the angle between them, while pairs of opposing boundaries are defined by their distance. The relevant notions of 'angle' and 'distance' were defined in an environment-centred coordinate system that relates only very indirectly to immediate body sensations. However, if the boundaries of an environment are vertical surfaces—for instance the walls of a room—then these measures can also be defined more directly. We have already seen that such boundaries can be defined 'positively' in terms of support affordances for the whole body, as well as 'negatively' as the limits of the navigable surface (see Section 2.3.5.2). These positive definitions can in some cases provide direct ways of defining relationships *between* boundaries, as well as of single boundaries by themselves.

This is particularly the case with angles between adjacent boundaries. For instance, the corner of a room is a place where two support affordances coincide: if the observer makes contact with two adjoining walls simultaneously, his locomotion movements are constrained in two separate directions, but in addition, there will be two separate support surfaces defined within his own body, where stable tactile contact is made, providing separate constraints on his balance system. The angle between the two walls can thus be given in body-centred terms, in relation to the angle between two surfaces within the observer's body. Note that although this is a body-centred representation, it maps directly to the environment-centred coordinate system, since it directly defines the constraints on locomotion in the corner being represented. Thus body-centred representations can provide useful independent definitions of the angles between boundaries in an environment. (Something similar is true for distances between opposing boundaries in an environment, in the special case where the agent can reach both opposing walls of an environment simultaneously.) The general principle that angles between surfaces can be defined in motor coordinates will be revisited in Section 2.9, when I discuss representations of objects as surfaces.

2.5.7 Summary: functions mapping from retinotopic to allocentric representations of boundaries in the environment

The operations discussed in this section are illustrated as a circuit diagram in Figure 2.5. The diagram depicts a set of functions that map from a retinotopic representation of the local environment to an allocentric representation of the environment and the observer's position and orientation within this environment.



'Imposed spatial structure'
allocentric location of external object (view cells)
allocentric boundary structure
allocentric observer location (place cells) / head direction
observer locating fn
retinotopic boundary structure
environment perception pathway
VISION

Figure 2.5: Circuitry involved in generating an allocentric representation of the environment, a map of places within the environment, and a representation of the observer's location and head direction within this map

- The **allocentric boundary structure generation function** transforms the retinotopic boundary structure into an allocentric boundary structure. (An **environment recognition function** attempts to map this to a stored environment representation in LTM; this function will be discussed in Section 2.16.1.1.)

- The **spatial structure function** imposes a spatial structure (topographical or topological) on the hippocampal system of place cells. This is shown as the green box, within which representations of the location of the observer and external objects are expressed.

- The **observer state identification function** maps the allocentric boundary structure onto a representation of the observer's current place within the system of place cells, and onto a representation of the observer's current head direction within the system of head direction cells.

- The function that generates representations of the allocentric location of external objects will be discussed in the next section.

An important idea in the architecture outlined here is that there are *two distinct perceptual pathways* within the environment/place system. One pathway computes an allocentric boundary structure from a retinotopic representation of the boundaries of the current environment, and uses this to impose a specific spatial structure on the map of places within the current environment. The other pathway computes representations *within* this map of places, detailing what is in the current environment at the current time (including both the observer himself and other external objects). This idea of two pathways will be reprised in my account of spatial representations of objects in Section 2.9.

2.6 Representing the location of objects in the current environment

As noted in Section 2.4.2.1, primates (including humans) can represent the location of external objects in the hippocampus in an environment-centred reference frame, as well as their own location. 'Place cells' in the hippocampus represent the location of the observer, while 'view cells' in the same area represent the location of a currently attended external object. In this section I will outline a model of how external objects are represented in an environment-centred reference frame, and how these representations relate to a representation of the observer.

2.6.1 Preliminaries: allocentric representations of self and other, and the concept of 'subject'

While place cells and view cells are in the same physical location in the hippocampus, there are several reasons why they should be modelled as occuring in two separate media. For one thing, very different computations are involved in identifying one's own location and that of an attended external object. (In fact, as I will show below, the computation required to identify the allocentric location of an external object takes the allocentric location of the observer *as an input*, so the medium holding the external object's location must be thought of as being further along a processing pathway than the medium holding the observer's location.) But just as importantly, it is essential that the observer can tell the difference between a representation of his own location and a representation of the location of an external object. A simple way of doing this is to activate these representations in different maps.

While the observer's location and the location of an external object should be represented in different maps, the maps are of the same environment: we can therefore envisage that the spatial structure imposed on the system of place cells (as described in Section 2.5.2) is also imposed on the map of view cells. These two maps are therefore separate, but isomorphic in structure. This means we can envisage a one-to-one mapping between representations of places in the two maps.

A final point is more subtle. Assume the system of view cells represents the location of 'the currently attended object'. This is normally taken to be an external object separate from the observer—specifically, the object currently being viewed. (This assumption is built into the term 'view cells'.) However, the observer can in some circumstances attend to *himself*. It is important that somewhere in the environment/place system there should be a medium that can represent the location of *either* the observer *or* an external object, at different times. For instance, in language, spatial predicates can apply equally to onesself or to external individuals: we do not want to have to assume two separate systems for representing such predicates, one for ourselves and one for others.

In several domains of cognition there is good evidence that the neural media that support representations of the self also serve to support representations of attended external agents. This idea is best attested in the domain of simple limb movements: the classic 'mirror system' hypothesis posits that areas of premotor cortex subserving the planning and control of simple motor gestures (e.g. movements of the arms/hands, legs, head and mouth) also serve to represent these gestures when they are perceived in other agents (see e.g. Rizzolatti *et al.*, 2001; Buccino *et al.*, 2001). There is similar evidence in the domain of emotion representations (see e.g. Singer *et al.*, 2004). The model I will propose will envisage that the medium representing the allocentric location of an attended external object can also hold the observer's own allocentric location, in the situation where the observer has attended to himself. I will use the neutral term **subject** to describe the individual whose location is represented in this medium. In some circumstances the subject is onesself; in others, it is an external individual. The location holding a representation of the currently attended individual will therefore be termed the **current subject location**.

In the current subject location medium, the problem of distinguishing between representations of one's own location and that of another individual is again raised—but in a form which admits of a solution. In the model I propose, there are two distinct ways of *creating* a representation in the current subject location medium: one for the location of the observer, the other for the location of an externally attended individual. Again, deictic representations play a central role: the way to identify whether the active location in this medium is that of the observer or of an external individual is to refer to the operation that created this representation.

2.6.2 The external object location function

External objects in the observer's environment will be represented on the observer's retina, if they are within his field of view. At any given time, the observer computes a retina-centred **saliency map** of interesting visual stimuli (Gottlieb *et al.*, 1998; Thompson and Bichot, 2005), and there are mechanisms for selecting the most salient stimulus wihthin this map (Itti and Koch, 2001). This is the stimulus that is classified by the object classification system (Moore and Armstrong, 2003; Zhang *et al.*, 2011). To create allocentric representations of the location of visually perceived objects, locations in the retinotopic saliency map must be converted to environment-centred coordinates.

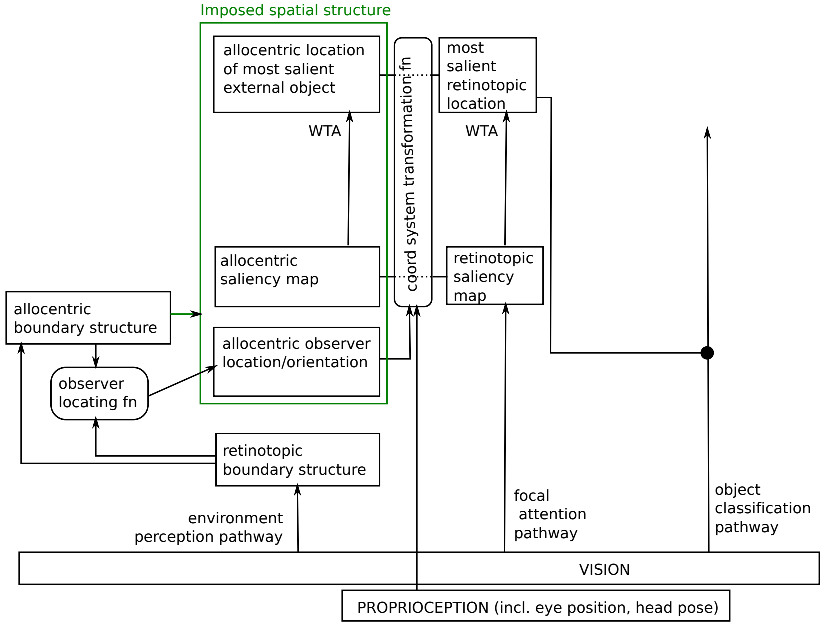
The object location function has to identify not only the position of an object in its environment but also its spatial relationship to the environment more generally: this involves an indication of its orientation in relation to the environment, and might also include an indication of what surfaces it is resting on or constrained by. These issues can only be considered when we have a model of object shape. I will introduce a model of object shape in Sections 2.9 and 2.10, and provide a fuller account of the spatial relationship between an object and its environment in Section 2.12; in the present section I will just consider how the observer identifies the *location* of external objects in his environment.

In the traditional account of how an observer generates an environment-centred representation of the location of a visually attended external object, a key requirement is that the observer knows his *own* location (and head direction) in the environment. Each time the observer moves in the environment, or changes his head direction, the representation of his own location and head direction will be updated. Given this assumption, the observer's retinal coordinate system can be mapped quite directly onto an environment-centred coordinate system: each point on the retina can be mapped to a line of points in the environment-centred representation, and if the distance of the object at the retinal point can be computed, this can be mapped to a specific position on this line.

There is some evidence that the saliency map is converted as a whole into allocentric coordinates, or at least into coordinates that are stable over eye movements. When subjects in a behavioural experiment are asked to count the dots on a page, they typically saccade to each dot while counting. In the classical model of this process, a saliency map represents the dots: at each iteration, a single dot is selected as 'the most salient', attended to, and inhibited, so that items in the map are progressively removed until there are none left (see again Itti and Koch, 2001). Given that subjects typically move their eyes at each iteration, the saliency map must be stored in a medium that is stable over eye moments. We must therefore envisage a function that converts the whole saliency map between retinotopic and allocentric coordinates. At the same time, we can envisage a function that converts the selected 'most salient item' from retinotopic to allocentric coordinates. This function is likely to be more accurate, since it can make use of a more accurate measure of distance, incorporating cues that can only be generated through object classification as well as cues like stereopsis that can be calculated in parallel. It is also likely to provide more information—in particular, information about the *orientation* of the attended object in its environment. This can only be delivered by a system that computes information about the object's shape, as noted above; and this can only be done for a single selected object at a time (see again Sections 2.9 and 2.10). Accordingly I will assume two functions, one converting all salient locations in parallel from retinal to allocentric coordinates, and one converting just the most salient location.

There are two final points to mention. First, I will assume that both these functions operate in both directions, converting allocentric representations to retinotopic ones as well as the other way round. Second, while the process of converting between retinotopic and allocentric coordinates is known to involve computation of several intermediate reference frames—namely frames centred on the head and body—I will omit these from the model, and refer to simplified functions that make the computation directly.

The operations discussed in this section are illustrated as a circuit diagram in Figure 2.6, which extends the diagram in Figure 2.5. There is now a focal attention pathway and an object classification pathway, as well as an environment perception pathway. (The most salient retinal location gates the input to the object classification pathway, so that classification resources are focussed on the selected salient object.)



Imposed spatial structure
allocentric location of most salient external object
WTA
allocentric saliency map
allocentric observer location/orientation
allocentric boundary structure
observer locating fn
retinotopic boundary structure
coord system transformation fn
most salient retinotopic location
WTA
retinotopic saliency map
environment perception pathway
focal attention pathway
object classification pathway
VISION
PROPRIOCEPTION (incl. eye position, head pose)

Figure 2.6: Circuitry involved in translating between retinotopic and allocentric representations of the location of external objects

Note that the location of place cells is now labelled as holding the 'allocentric location of the most salient external object'. Note also that I assume a winner-take-all (WTA) operation is implemented in for the saliency map in both retinotopic and allocentric coordinates. The latter operation will be discussed more in Section **??**. Finally, note that while the coordination transformation function is expressed as a single function, it nonetheless takes input from proprioceptive representations (of the angle of the eyes in the head, and of the head in relation to the body).

2.6.3 Representing onesself as an attended object

A final piece of circuitry allows the system of view cells that represent 'attended objects' to represent the observer himself, as envisioned in Section 2.6.1. A key element in this circuitry is an operation that selects between two alternative cognitive modes. In **external perception mode**, the observer configures the environment/place system to represent the location of an external object. In **self perception mode**, the observer configures this same system to represent his own location. In the former mode, the attended allocentric location is chosen to be the most salient external location, as described in Section 2.6.2. In the latter mode, the attended allocentric location is copied directly from the medium representing the current observer location. (Recall from Section 2.6.1 that there is a one-to-one correspondence between locations in these two media.) Since the medium representing the attended allocentric location can represent either the location of the observer or of an external agent, I will now refer to it as holding the **allocentric subject location**. To determine whether the location in this medium is that of the observer or that of an external object, we must refer to the operation which established one or other of these modes of connectivity.

Clearly, these two modes must be alternatives to one another. We must envisage competition between them, and different sources of evidence favouring each mode. External perception mode should be favoured if there are salient external stimuli. I propose that the trigger for this mode is activated by summed activity within the saliency map. Self perception mode can be activated by two separate kinds of evidence. Firstly, a decision to execute an action should initiate this mode. When the observer executes an action, he must represent his own position in the environment in the same medium that represents the actions of other agents. The decision to initiate an action comes from the action planning system, which will be discussed in more detail in Section **??**. Secondly, if the observer is in the process of perceiving an action happening externally to him, which he did not initiate, it is still possible that he is a participant in this action. There are several ways in which he can realise that he is a passive participant in an external action. Most concretely he can feel his body being influenced by an external source. This sensation will be registered by his somatosensory system. If the action is carried out by an animate agent, it is likely this agent will fixate him visually, so another indication that he is the target of an action is if he experiences the sensation of direct eye contact after attending to another agent. Perceiving direct gaze from another agent is qualitiatively different from perceiving gaze directed elsewhere: it is strongly represented in several specialised brain areas (see e.g. Jellema *et al.*, 2000) and has special effects on the observer's level of arousal (see e.g. Kleinke and Pohlen, 1971). There are several other consequences of activating self perception mode. One of these is an automatic activation of the representation of the observer himself ('me') within the object classification system. (This method of activating a representation of onesself is very different from visual classification, though of course we can also recognise a visual image of ourselves, for instance in a picture, or reflected in a mirror.)

The operations discussed in this section are illustrated as a circuit diagram in Figure 2.7, which extends the diagram in Figure 2.6. In this figure, self perception mode and external perception mode are control operations, that are mutually inhibitory (as illustrated by the link with the flat ends). External perception mode is activated by summed activity in the saliency map, while self perception mode is activated by somatosensory stimuli (unexpected touch sensations), signals from the planning system (specifically, the decision to act), and the sensation of a direct gaze originating in the visual object classification pathway. (This latter sensation is computed separately from object categories, as shown by the separate path.) Self perception mode gates open a link which copies the allocentric observer location representation into the allocentric subject location. External perception mode gates open a link which chooses the most salient location in the allocentric saliency map. Finally, self perception mode also activates a representation of the self (‘me’) in the medium holding the current object representation.

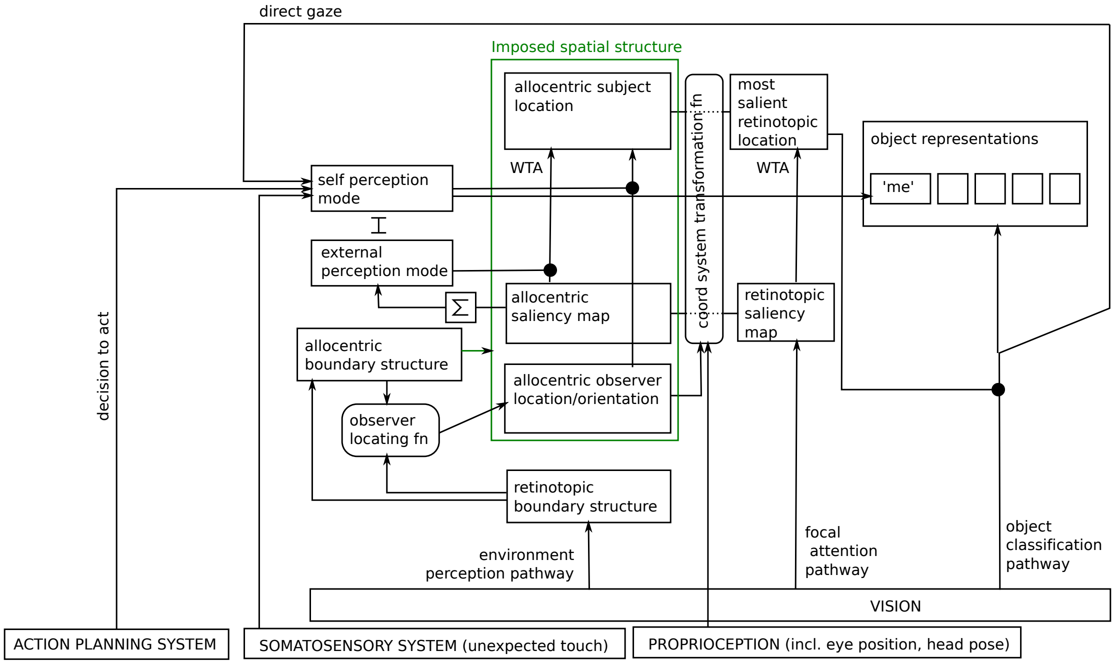

direct gaze
Imposed spatial structure
allocentric subject location
coord system transformation fn
most salient retinotopic location
object representations
'me'
WTA
WTA
self perception mode
external perception mode
Σ
allocentric saliency map
retinotopic saliency map
decision to act
allocentric boundary structure
allocentric observer location/orientation
observer locating fn
retinotopic boundary structure
environment perception pathway
focal attention pathway
object classification pathway
VISION
ACTION PLANNING SYSTEM
SOMATOSENSORY SYSTEM (unexpected touch)
PROPRIOCEPTION (incl. eye position, head pose)

Figure 2.7: Circuitry involved in selecting the self or another agent as the current subject

2.7 Representing environments and object locations in LTM

As already noted in Section 2.4.1, the observer stores representations of the shapes of the environments he encounters in LTM, probably in the parahippocampal place area. Now that we have a model of how environments are represented perceptually and in working memory (Section **??**) and of how the locations of objects within environments are stored in working memory (see Section 2.6), we can consider how environments are represented in LTM, and how the locations of objects in these environments are stored in LTM. I will consider the first question in Section 2.19.1, and the second question in Section 2.7.2.

2.7.1 LTM environments

2.7.1.1 Theoretical background: Damasio's model of convergence zones

The theoretical framework for the model of LTM that I will sketch is the model of 'convergence zones' introduced by Damasio and Damasio (1994). In this account, LTM representations of some stimulus (say an object or an environment) is a pattern of neural activation in some high-level association cortex, that has rich connections to cortices that encode the representations of the stimulus generated in a range of sensory and/or motor modalities while the observer is actually engaging with it. The association cortex is called a **convergence zone**, because of its rich connections with these sensory and motor areas. Damasio and Damasio's central idea is that a stimulus is represented in LTM as an arbitrarily selected pattern of activation within a convergence zone. The content of the LTM representation is defined by the way this arbitrary pattern of neurons is connected to sensory representations: it holds 'bindings' between sensory representations in different modalities. For instance, the LTM representation of a particular dog could be implemented as a pattern of activation in a convergence zone that is separately linked to visual representations of the dog, to auditory representations, to tactile representations and so on.

In this scheme—as in most models of LTM—the representation of a given stimulus in LTM does not involve any particular pattern of neural activation; rather it takes the form of a *potential* for activation of a particular pattern of activity, defined in a pattern of synaptic weights. It is only when an LTM unit becomes active that this potential is actualised. Importantly, when it is actualised, it produces a pattern of activity in the sensorimotor regions which is similar to that produced by direct perceptual or motor *experience* with the stimulus in question. So activating a LTM representation of the stimulus is like reliving the experience, in some sense.

2.7.1.2 Storing the shape of an environment in LTM

I will assume that there is a convergence zone dedicated to storing environments and their spatial structure, and I will term the arbitrary patterns of activation in this neural medium **LTM environments**. As already discussed, the parahippocampal place area plays an important role in encoding both perceptual representations of the observer's *current* environment, and in representing *remembered* environments. For simplicity's sake, I will imagine two areas of the PPA, one functioning as a convergence zone, and one holding perceptual representations of the shape of environments.

Recall from Section 2.5.2 that when the observer encounters a new environment and navigates around it, he constructs a perceptual representation of his current environment in the PPA. This representation is holistic, and allocentric: it is an allocentric boundary structure. I now propose that to represent this new environment in LTM, the observer creates a new LTM environment representation in his PPA convergence zone—i.e. a new random pattern of activation in this zone—and links this pattern to the currently active allocentric boundary representation. After this link is created, if the allocentric boundary structure generated through perception of the *current* environment is close enough to one that's associated with a LTM environment, this LTM environment will become activated, and we can speak of the observer 'recognising' his current environment as one he has encountered before. I will consider this scenario in Section 2.7.1.3. But we can also envisage that the observer activates an LTM environment simply because he is recalling a distant environment that he encountered some time before. In this case he is simply *remembering* a previously encountered environment. I will discuss this situation in Section 2.7.2.2.

2.7.1.3 Recognising a perceived environment

2.7.2 Representing object location in LTM

2.7.2.1 LTM individuals

Alongside the convergence zone that stores LTM environments, I will assume there is a second convergence zone dedicated to storing representations of token individuals, or **LTM individuals**. Again, there is good evidence that LTM representations of individuals are stored in parahippocampal areas; see e.g. (...). I will have a lot more to say about LTM individuals in Section 2.11. For the moment, I am just concerned with how individuals can be represented in LTM as occupying particular locations in a remembered environment. For instance, how can the observere encode in LTM the fact that his desk occupies a certain place in his office?

2.7.2.2 Interpreting the map of places in a remembered environment

Consider a scenario in which the observer re-activates an LTM environment in a memory context. Presumably in this scenario he must somehow disengage the mechanism that delivers an allocentric boundary representation from perception, and enter some kind of special 'memory mode', in which sensory and motor representations are generated through their links to LTM representations in convergence zones rather than through current perceptual and motor experiences (see e.g. ??? for evidence that such cognitive modes exist).

It is instructive to consider what happens to the hippocampal map of places in this memory context. The interesting thing to note is that the currently active allocentric boundary representation imposes a spatial structure on the system of place cells that represents a map of places in the current environment. In this memory context, 'the current environment' is not the environment currently being *perceived*, but the one currently being *remembered*. This means that when the observer activates an LTM environment representation in a memory context, he also generates a map of the spatial locations in this environment, with the same topological and/or topographical structure as the map that would be generated during direct experience with this environment. This opens the way for a simple account of how the observer remembers the locations of objects in his environment, which I will introduce in Section 2.7.2.3.

2.7.2.3 Linking LTM individuals to points in the allocentric map of places

My basic proposal is that memory for object locations is stored in a network that links representations of place within the hippocampal map of places bidirectionally to LTM individuals, in a way that is modulated by the currently active LTM environment. In this proposal, the dynamics of the memory system are as follows. First, an LTM environment is activated. This modulates the links between LTM individuals and places so that a particular set of links are in force. It also activates an allocentric boundary structure holistically representing the shape of the remembered environment. This in turn imposes a spatial structure on the hippocampal system of places. At this point—but not before—place representations have meaningful interpretations, and the mapping between LTM individuals and places can be interpreted.

In practical terms, if the observer recognises his current environment, he will also activate, from LTM, representations of the remembered locations of objects in this environment: for instance, the location of his desk in his office.

I will return to the issue of LTM representations of objects and environments in Section **??** and 2.12, after introducing a model of the spatial representations of objects.

2.8 Representations of local relationships between environments in LTM and the sensorimotor system

2.8.1 Relationships between adjacent environments

There are typically places in an environment where the observer can enter another environment. For instance, if my current environment is my office, the door which leads into the corridor is one such place: it links the office and corridor environments. I will use the term **adjacent environments** to describe environments linked this way.

When the observer steps through the door of the office into the corridor, he must establish a new environment, with a new shape, whose map of locations therefore has a new spatial structure. And he must reposition and reorient himself within this new spatial structure. These processes involve work both within the sensorimotor system and within the LTM system. In the LTM system, the observer must activate a new LTM environment: either a brand new one, if he is going through the door for the first time, or an existing one, if he has expectations about the environment he will find when he is through the door. (I will focus on the latter case, so I can describe a fully learned representation of adjacent environments.)

The basic circuit within the LTM system is a recurrent one: when the agent goes through the door, a new unit is selected as the 'current LTM environment'. The new unit is associated with its own allocentric boundary structure, and therefore imposes a new spatial structure on the map of places. The operation that triggers this update involves arriving at a particular place in the first environment, establishing a particular orientation, and performing a particular action, that I'll call a **reconfiguration action**. (In this case, it is simply the action of walking forward through the doorway, but I will discuss more complex cases in Sections 2.9.6.1 and **??**.) When this operation is performed, a new current environment is established, with its own allocentric boundary structure: this imposes a new spatial structure on the map of places. Within this new structure, the observer has to update his own location and orientation. So the update operation not only establishes a new LTM environment: it also activates a new place, and a new orientation.

This is ordinary indexing: there's a function that maps from the current LTM environment, plus a place within this environment (somewhere on one of its boundaries), plus an orientation, plus a reconfiguration action (of which I'll say more in Sections 2.9.6.1 and **??**), delivers a new current LTM environment, plus a change to the scaling factor, plus a change in orientation.

One important idea in here: I'll augment the current model of environments with two types of *navigation goal*: goal places, and goal local environments.

2.8.1.1 Goal places

The simplest type of navigation goal is a **goal place**. The goal place is represented in a medium that's isomorphic to the current observer place and current subject place.[9] The goal place, and the current observer place together provide input to a function that generates navigation actions through the observer's current environment. These will be discussed in much more detail in Chapter **??**; for now I just want to use them to express operations that transition between environments. I assume there's a medium representing the *candidate* goal places within the current environment, and another representing a *selected* goal place from this set of candidates.

2.8.1.2 Adjacent goal environments

A second navigation goal is an **adjacent goal environment**. This represents an enviroment immediately adjacent to the current environment. As before, I envisage a medium holding a set of candidate adjacent goal environments and another medium holding a single item selected from this set. The candidate adjacent goal environments represent all the environments that the observer can get to directly from his current environment.

I envisage a function called the **adjacent environment transition function**, which maps the current environment and a selected adjacent goal environment onto a goal place, an orientation, and a transition action. This function tells the observer where to go in his

[9] I think that it should be called the 'goal *subject* place', but I'll stay away from this issue for the moment.

current environment if he wants to get into a selected adjacent environment, and what orientation to adopt and transition action to perform to enter this environment when he gets there. So the environment transition function is actually represented in goal-based terms: if you reach a given place, assume a given orientation and do a given transition action, then the selected *goal* adjacent environment will become the new *current* environment. (And certain updates to the current orientation and scaling factor will be applied.)

2.8.2 Hierarchical relationships between environments

Adjacency relations between environments define a topological structure: a graph. However, we do not represent the spatial world as a single giant graph of low-level environments: rather we represent the topological structure of selected local regions of the world, and then recursively represent the topological structure of these regions within larger local regions and so on. For instance, the environments that neighbour or are close to my office are represented as environments within a particular building (the Owheo building). The Owheo building itself is represented as an environment within a larger region—say the campus of Otago university—and this in turn lives inside a larger environent (say the city of Dunedin), and so on up. The observer must be able to represent his current environment—say my office—but also to represent the **super-environment** within which this environment is situated. He must also be able to transition between hierarchical scales: for instance, when the observer leaves the Owheo building, his super-environment stops being the Owheo building and starts being the Otago campus.

Evidence for this hierarchical model of environments comes from a classic experiment by Hirtle and Jonides (1985). Subjects were asked to enumerate landmarks in a university campus. An analysis of the sequential structure in this free recall task showed a clustering of landmarks. In a followup experiment, subjects were asked to estimate distances between landmarks. It was found that estimates were greater for landmarks from different clusters than for landmarks from the same cluster, even when the actual distances were roughly the same. This is good evidence that environments are hierarchically organised into super-environments in the kind of way I am proposing.

2.8.2.1 The concept of a super-environment

What defines a super-environment? It is partly defined topologically, as a set of environments that are highly interconnected: for instance, there are typically many ways of getting from any one environment in the Owheo building to any other environment. But it can be defined in other ways too. For instance, from a distance, the Owheo building is clearly a single object: it has the same kinds of property that at a smaller scale define manipulable objects (convexity, continuity, textural homogeneity). From within, the environments in the Owheo building have a lot in common: the floors, walls and ceilings look similar. Functionally, the environments in the Owheo building also have similarities: they are used for similar purposes; partly for this reason, agents using the building frequently move from one of its environments to another. For all of these reasons, when the observer gains experience of this building, it makes sense for him to represent it as a single unit.

2.8.2.2 A circuit defining the spatial structure of a super-environment

[A key recursive idea for the future here: you never just represent an object: you're always obliged to *also* represent that object's local environment. I hope this will be relevant for existential sentences and locative PPs. When I direct focal attention to a location, I'm identifying the object at this location in one perceptual pathway, and I'm identifying the environment at this location in another, parallel perceptual pathway—and I'm not done until I've found representations in each pathway that are consistent with one another (in the sense that there's a recognised spatial relation like support or containment linking the two representations).]

I propose that alongside the medium representing the current LTM environment, there is a medium representing the **current LTM super-environment**. These media are parallel: each environment in the observer's model of the world is represented in both media, and there are 1:1 links between the media, so that any environment established as the current LTM environment can be re-established as the current LTM super-environment and vice-versa. In addition, I assume that each environment representation E_i in the current LTM environment medium is linked to exactly one *different* environment representation in the current LTM super-environment medium, that identifies the super-environment to which E_i belongs.

The LTM super-environment activates a particuar set of adjacency relationships between its component LTM environments. These are expressed within the account of adjacency given in Section 2.8.1.2: the currently active LTM super-environment provides additional input to the adjacent environment transition function defined in that section, that specify how to enter a given adjacent goal environment.

As mentioned in Section **??**, I suggest that the LTM super-environment also defines spatial relationships between its constituent LTM environments that are *not* adjacent, that permit the observer to navigate between arbitrary pairs of constituent environments. For instance, a representation of the Owheo building as a super-environment might indicate to the observer how to get from my office to the downstairs lobby, or from the secretary's office to the teaching lab. These route specifications are important because of the functional characterisation of a super-environment: the tasks that the observer performs frequently call for him to travel between many pairs of non-adjacent environments in the Owheo building, so it is worth storing these routes explicitly, rather than relying on some general graph-search algorithm.

The idea that the observer stores relationships between distant environments within a super-environment using precompiled sequences of intermediate environments receives support from several directions. Evidence that 'well-learned routes' are stored in specific ways in the brain comes from Hartley *et al* (2003); these authors present fMRI data showing that the caudate nucleus is more activated when subjects move along a well learned route than when they are navigating a new route created from general knowledge of the spatial layout of their immediate environment. This finding echoes a similar finding in rats (see e.g. Packard and McGaugh, 1996). Well-learned routes appear to be represented as stereotypical action sequences: they are established slowly, using reinforcement, and once learned, enable fast, automatic behaviour. There is also recent evidence that the cerebellum is involved in encoding spatial knowledge used for navigation; see e.g. Rochefort *et al.* (2013). The cerebellum is good at storing stereotypical behavioural sequences (see e.g. Penhune and Steele, 2012).

In the model I propose, stored routes between arbitrary environments within a super-environment are also expressed using the terminology of navigation goals. I assume that the currently active LTM super-environment indicates not only the set of *adjacent* goal environments, that immediately adjoin the current environment, but also a set of more distant goal environments that can only be reached by moving through *sequences* of adjacent environments. These more distant goal environments are still within the currently active super-environment, so I will call them *local* goal environments. Specifically, I will assume that the currently active LTM super-environment is directly linked to a set of **candidate local goal environments**, from which a single **selected local goal environment** can be chosen. Each selected local goal environment is linked to a *sequence* of adjacent environments within the current super-environment that connects the current environment to this selected environment. [10] The environments in this sequence can be adopted one by one as the selected adjacent goal environment (see Section 2.8.1.2), to enable navigation to the local goal environment.

To implement the above idea, I introduce another function, analogous to the adjacent environment transition function introduced in Section 2.8.1.2, called the **local environment transition function**, that maps the current environment and a selected local goal environment to a sequence of adjacent environments. Like the adjacent environment transition function, this function is conditioned on the currently active LTM super-environment, so it represents extended sequential paths within this super-environment.

In summary, the currently active LTM super-environment determines three things. Firstly, it directly identifies the component environments within it, through direct links to a set of candidate local goal environments. Secondly, it determines the adjacency relationships between these component environments, through the adjacent environment transition function. Finally, it determines a set of extended paths involving these component environments, through the local environment transition function.

2.8.2.3 Transitions between environments and super-environments

If the observer is navigating within the Owheo building, his current LTM environment will provide a representation of the spatial structure of the particular room or corridor he is currently in, and his current LTM super-environment will providse a representation of the structure of rooms and corridors etc within the building. But if the observer leaves the Owheo building, a new LTM super-environment must be identified—say the Otago

[10] Or perhaps more accurately, to a set of possible sequences, from which a winner must be chosen. I assume that the winning sequence is the one associated with least effort, or something like that.

University campus—in which the Owheo building is a component environment. If he goes back into the Owheo building, the Owheo building must again become the super-environment.

In the scheme I propose, these transition operations are accomplished via two related mechanisms. One is implemented in the links between the medium representing the current LTM environment and the medium representing the current LTM super-environment. I assume that

As discussed in Section 2.8.2.2, each environment in the observer's model of the world is represented in both media, and there are direct links from the unit representing

2.8.2.3.1 Boundary environments and hierarchical transition operations [I don't think this section is well placed. . .]

We have to imagine that there are special *boundary environments* within a super-environment whose adjacent environments are in a new super-environment. For instance, in the super-environment of the Owheo building, the ground floor lobby is a constituent environment that is also a boundary environment, in the sense that there is a place within it (the entrance to the building) at which the subject can *leave* the building, and arrive at a place in a wider super-environment (say Union Street).

The transition operations that allow the subject to leave the current super-environment are defined within the super-environment, in the same way as normal adjacency relations: they are triggered within a specified boundary environment (e.g. the ground floor lobby of the Owheo building), when a given place is reached (e.g. the door of the building) and a given orientation is established (e.g. facing the door) and a particular transition action is executed (e.g. walking forwards). The difference is that the transition establishes a new super-environment, and a new place and orientation within this super-environment.

Here it is useful to consider two alternative scenarios. In one, when exiting the Owheo building, the observer directly finds himself at a place within the super-environment. For instance, imagine the Owheo building is situated within a large open space, like a field, which can be freely navigated (within its boundaries). The building is represented as being 'at' a place within the field environment, and when leaving the building, the observer is represented as having a particular orientation and place in this same environment. In another scenario, when exiting the Owheo building, the observer immediately enters an environment that neighbours the Owheo building within its super-environment. For instance, if when exiting the Owheo building the observer finds himself at a certain position on a road, he could represent the road as the super-environment of the building (in which case the building is 'on' the road) or he could represent the building and the road as adjacent environments within a wider super-environment (e.g. the campus). In this latter case, there is an asymmetry between these adjacent environments, because they are in different super-environments: the road is in the campus super-environment, but the ground floor lobby is in the Owheo building super-environment.

In either case, when the observer has left the Owheo building, this building still needs to be represented, but in a quite different way: by its external surfaces rather than its interior structure. Its external surfaces now form boundaries for navigation actions: the subject can walk along one of its walls, or round the whole building, or possibly even climb on top of it. The observer is now well-placed to represent the three-dimensional shape of the building, as a collection of *external* surfaces. I will discuss this type of representation in much more detail in Sections 2.9–2.12.[11]

along the wall of the building: while from the inside, the building has internal complexity, from the outside, this wall just provides a boundary for the locomotion environment associated with the street.

Important somewhere here to mention the scaling factor. When you transition to a super-environment, you change scale: the map of places is represented more coarsely. This change in the scaling factor has to be represented.

the street outside the Owheo building is

2.8.2.3.2 The local environment topology function as a map of places And now for the key idea: the map of places is structurally analogous to the local environment topology function.

In my model, there are two active environment representations at any time. Say I'm in my office in the Owheo building. My current environment is my office. My current super-environment is the structure that determines

2.8.2.3.3 [What you must *absolutely* make sure to do is to connect this transition to the transition in which an *object* is re-represented as an *environment* (see Section **??**). Remember, there's a 1:1 mapping between objects and environments.]

2.8.3 Parallel representations of sub-environments

I suggest that there can be several saliency maps active at any given time, representing maps of places within different sub-parts of an environment. For instance, in a room, there are several sub-environments: the floor, different walls. I suggest that the saliency map associated with the room is in fact *several distinct* saliency maps, one associated with each of these sub-enviroments. Nonetheless, there is competition to find the single most salient location right across these maps. The winning location identifies a place *in a particular map*: it's not just the place that wins, but the map as well. So the salient location has a function in choosing the currently active sub-environment.

My guess is that this process is described in existential sentences. When I say *There was a picture on the wall*, my suggestion is that *There* identifies a salient location, *a picture* identifies the object that's at this location, and *the wall* identifies the sub-environment

[11] An interesting note: this shape-based representation of the building allows me to define a metrical representation of places within it, which I suggest lives alongside the topological representation of corridors, offices, lobbies and so on. (Recall from Section **??** that the map of places within an environment is structured using a mixture of topological and topographical representations.) This can help the observer to explore the building, or to find new routes between its component sub-environments.

of the current environment in which the location is represented, and *on* describes the relationship between the sub-environment and the spatial structure of the identified object.

In this scheme, the idea is that shifting spatial attention involves both selection of a new object and selection of a new sub-environment: you can't have one without the other.

Where are these parallel saliency maps? Ultimately, I suggest they're in the motor system: one map for each separate effector-based motor system (e.g. one for each hand/arm, and within each hand/arm there are others for fingers, thumb etc).

2.9 Haptic representations of the shape of manipulable objects

In this section I will outline a model of how an observer represents objects in his environment that are small enough to be manipulable. This will not be a general model of how objects are represented—it will just provide one component of the general model. In the general model, we have to allow that environments that the observer can navigate in can also be objects in their own right: for instance, my house is a navigable environment, but it is also an object at a particular location within my street. In Section **??** I will propose a basic duality between objects and environments—that all objects can be re-represented as environments basic duality between objects and environments and vice-versa. But this account will hinge on a general account of transitions between environments; the account given in the current section, of smallish, manipulable objects, will contribute to this account of transitions.

It is a well-acccepted idea that manipulable objects are represented by their motor affordances in some way, and that these representations are held in the visuomotor pathway running through parietal and premotor cortex, with contributions from somatosensory cortex. My main proposal in this section is that this visuomotor pathway should be thought of as an integral part of the environment/place system—that is, that the representations computed in this pathway conform to the same general scheme as those computed in the pathway representing the observer's navigable local environment and places within it.

2.9.1 Overview

The basic idea I will propose is that manipulable objects can be thought of as environments, which are analogous to the environments in which the observer moves—except objects are environments that are navigated in by the observer's *effectors* rather than the observer in his entirety. If a manipulable object is relatively large, the relevant effectors are complete arm/hand systems; if it is small, the relevant effectors may be the fingers within a single hand. I will propose that the most basic representations of objects construed this way is the haptic system.

I will use two example objects to discuss this idea: a large dining room table and a small cup. A table is a useful reference object, because something this size is at the limit of what can be considered 'manipulable' (at least by a single person); the surface of a table is also something whose geometry is easy to compare to that of a locomotion environment like a room. A cup is a more prototypically sized manipulable object, and has more complex geometry.

The haptic effector I will focus on in both cases is the hand. But since the example objects are of different sizes, somewhat different motor systems must be envisaged to move the hand around the object in the required ways. The motor movements that explore a table haptically are reasonably large movements of the arm (or arms), and possibly movements of the trunk, and possibly also locomotion movements. There are two relevant motor systems in this case: one is the parietal/premotor system that represents surfaces in the observer's perispace as motor maps of hand/arm states (discussed in Section 2.3.5.1); the other is the locomotion system discussed in Section 2.3.5.2. The motor movements that explore a cup haptically are movements of the arm (or arms) that transport the hand small distances, and movements of the wrist and fingers. The relevant motor systems here are those that represent the observer's perispace as in motor maps of hand/arm states.

The observer has to be able to compute two representations of the object. Firstly, he must compute a representation of the *shape* of the object. We can specify two characteristics of this representation already. For one thing it is a three-dimensional measure. I have already considered mechanisms for computing the 'shape' of a locomotion environment (see Section 2.4, but these were for computing a two-dimensional measure of shape, at least at a first approximation, since the surfaces that humans normally navigate in are constrained to be roughly horizontal. There is no such constraint for haptic navigation, since the hands can be moved vertically as well as horizontally: so the shape representations supporting this type of navigation will be more complex. For another thing, the computed representation of the shape of an object should be given in a coordinate system centred on the object: it should not make assumptions about the location of the observer in relation to the object, since this can vary. So a mapping must be implemented between body-centred haptic representations of object shape and allocentric object-centred representations.

Secondly, the observer must compute a representation of the object's geometry in relation to its own environment. The observer must represent the *location* of the object in its environment. He must also be able to represent its *orientation* in its environment. (There is a difference between an upside-down table and a table the right way up, and a difference between a table oriented parallel to the walls of a room and one oriented diagonally.) Again, these tasks have their own complexities. For one thing, the object to be localised could be in *the same environment as the observer*, or in *a separate environment*. The table and cup are helpful in illustrating this distinction. A table is something that is likely to rest on the same horizontal surface that the observer locomotes across (for instance the floor of a room). But a cup is something that is more likely to be found on some *other* surface. (For instance on a table!) If the observer wants to compute a spatial representation of a cup on a table, he must represent the position and orientation of the cup in relation to the table, not in relation to his own environment. Clearly there is still a sense in which the cup is also 'in' the observer's environment: in Section **??** I will introduce a model in which environments can be nested within each other that will account for this. For current purposes the main point is that an object's own most immediate environment might not be the one that the observer is in: so the observer must be able to represent the location and orientation of an object in relation to an environment that he is *viewing*, rather than occupying.

The two desiderata summarised above are related, in that representing the orientation of an object in relation to its environment must make reference to the object-centred frame of reference used to define its shape. So we should begin by considering how to represent the shape of objects in allocentric terms, and then consider how to represent the location and orientation of objects in their environments.

2.9.2 Representing object shape: a general proposal

I propose that the shape of an object is represented in exactly the same format as a navigation environment: as an allocentric boundary structure, defined as a set of pairs of opposing and adjacent surfaces. I will first summarise how the observer uses allocentric boundary structures to represent the shape of his local environment, then summarise how a similar scheme can be used to represent the shape of manipulable objects.

2.9.2.1 Recap: boundary structures for representing the observer's local environment

In the model introduced in Section 2.5 the parahippocampal place area computes a representation of the boundaries between the surfaces of his current environment in a retinotopic reference frame, and converts this representation into an allocentric reference frame, termed an *allocentric boundary structure*. This allocentric structure then imposes a 'spatial structure' on the medium holding a map of places and orientations in the current environment, that stores the location and orientation of the observer himself, and (separately) of an attended 'subject' (that can be an external individual or the observer himself). In a separate pathway, the retinal boundary structure is used to compute a representation within this medium of the current location and orientation of the observer.

2.9.2.2 Boundary structures for representing the shape of manipulable objects

I suggest that exactly the same scheme is used to represent the 3D shape of a manipulable object. Specfically, I propose that an object's shape is first represented as a structure of surfaces defined in observer-centred reference frames, comprising pairs of adjacent and opposing surfaces. (These structures are learned first in motor coordinates in somatosensory cortex, and later in equally observer-centred retinotopic coordinates.) This observer-centred boundary structure is then converted to an allocentric boundary structure, centred on the object, plus a representation of the object's orientation in relation to the observer (which is used in turn to generate a representation of the object's orientation in relation to its own environment). The object-centred boundary structure then imposes a 'spatial structure' on a medium holding an object-centred map of places in, on, or near the object. In a separate pathway, the observer-centred boundary structure provides input to a function that computes the places occupied by the observer's own effectors in this map medium.

In the remainder of this section I will discuss the stages in this pathway one by one.

2.9.3 The atomic components of shape representations: somatosensory representations of local shape

I have already discussed representations of surfaces generated in the haptic and visual systems: see Sections 2.3.2, 2.3.3 and 2.3.5.1. In those sections I described representations of general *types* of surface, with different orientations, peripersonal locations and curvatures. I suggest that these representations constitute a set of primitives from which particular shape representations are composed.

I will begin in Section 2.9.4 by discussing how the individual boundaries of a surface are represented. In Section **??** I will consider how representations of the boundaries of a single surface can be combined to derive a representation of the shape of this surface. In Section 2.9.6 I will consider how representations of the geometry of pairs of surfaces are computed.

2.9.4 Defining the individual boundaries of a surface

If a shape is composed of more than one surface, then it is important to define where one surface finishes and the next one begins: these points constitute a *boundary* between the two surfaces.

The easiest case to think about is the boundary where two flat surfaces meet, which is an **edge**. As usual, I will define this kind of boundary first within the haptic modality. Assume that the observer's hand is travelling along a flat surface, making contact at the surface formed by the palm, finger and thumb tips. In haptic terms, there are two basic types of boundary. An adjoining surface that rises 'up' (in the coordinate system of the hand) is a barrier to the hand's navigation: the hand cannot navigate on the surface beyond this boundary. An adjoining surface that drops 'down' creates a region where the hand cannot navigate without losing losing its stable support signal. Note that these definitions are very similar to the definitions of boundaries within the observer's locomotion environment, as discussed in Section 2.3.5.2: the notion of a bounded surface as permitting navigation with maintenance of stable support is key in both cases.

In fact, the same notion permits a useful definition of boundaries for curved surfaces. I will discuss this definition in Section 2.9.6.1.2. But first I will consider in a little more detail how discontinuous boundaries are represented in somatosensory terms.

2.9.4.1 Somatosensory representations of barriers

A barrier is represented as simultaneous contact within *two* surfaces of the hand. The angle of the barrier is given by the angle between these two surfaces. In fact we should be more precise: a barrier is a *concave* boundary, in the sense that the hand is 'inside' the space defined by the two external surfaces. The surfaces on the hand are, accordingly, *concave.*

The angle between the two external surfaces is given directly by the angle between the two surfaces on the hand. One component of this angle can be specified directly in terms of hand geometry: if we assume a standard reference pose of the hand, there is a fixed angle between each pair of contact surfaces in the hand. Another component depends on the current pose of the hand.

There is good evidence that somatosensory cortex computes representations of contact within two separate surfaces in the hand using a system of this kind. For instance, as already discussed, Hamada and Suzuki (2005) found representations of simultaneous contact by the index finger and thumb which varied when the angle between these digits was changed.[12]

2.9.4.2 Somatosensory representations of edges

The 'edge' of a surface can be haptically represented in two ways. Firstly, it can be represented in a signal indicating contact within *just one part* of a single hand surface. To identify that a surface within the hand is on the edge of a flat external surface, it is important that *part* of this surface is in a state of stable contact, but equally important that some *other* part *of that same surface* is *not* in a stable contact state. In this state, the agent can be sure that the contact surface drops away at the point between these two surfaces in the hand.

There is suggestive evidence that macaque somatosensory cortex computes edge representations of this kind. Fitzgerald *et al.* (2006) investigated the responses of SII neurons to surfaces contacting the twelve soft pads of the animal's fingers, when the fingers were aligned in a flat plane. Some of these cells were excited by contact on some fingerpads, but inhibited by contact on others. These cells have exactly the right properties to identify the edge of an external surface by touch, because they fire maximally when there is contact at one point in the surface formed by the fingers, and no contact at another point in that same surface (see the discussion in Haggard, 2006).

If the observer's fingers are hanging over the edge of a surface, an active movement is needed to identify the angle of the adjoining surface. The simplest action is to move the hand 'forward' (within its own coordinate system) and bend the fingers 'down'. Importantly, this action does not create a curved surface in the hand: it creates two separate planes within the hand. If the hand is configured so that stable contact is achieved within both these planes simultaneously, we have a situation analogous to contact with a barrier, except that the two surfaces in the hand now form a 'concave' space, with something 'convex' inside it.

[12] Get another reference too.

2.9.4.3 Stable contact with a discrete boundary

In all the above cases, a very important idea is that the boundary between two flat surfaces defines a stable support state *of its own*, with its own associated navigation routines: there are special ways of moving the hand that *maintain* the current haptic stimulus, whether this involves contact within a particular sub-part of a suport surface or contact with two support surfaces. Navigation in these cases is constrained separately by the two external surfaces, so haptic navigation movements can only move the hand along a single line.[13] But this line is very useful, for two reasons. Firstly, it allows us to compute the angle between two adjoining surfaces *for the surfaces as a whole*, rather than just for individual points on the boundary between them. Secondly, the line is defined within both surfaces: this helps define actions that *transition* the hand from one surface to the adjoining one. I will talk more about such actions in Section 2.9.6.1. But we first need to discuss how the observer can represent the shape of the surface that his hand is *currently* exploring. I will consider this in Section 2.9.5.

2.9.5 Representing the geometry of a single surface

The observer must have a way of distinguishing between surfaces with different shapes. Again it is helpful to consider a simple scenario where the surface currently being navigated by the hand is completely flat, and its shape is defined by flat boundaries that form straight lines onto the navigation surface. There is nonetheless an important difference between a square surface, a triangular surface, a rectangular surface, a surface with a circular perimeter and so on.

As already discussed in Section 2.1, I want to argue that the problem of representing the shape of objects is analogous to the problem of representing the shape of navigation environments. In this section I will illustrate this idea in the case where the analogy is closest, and argue that the shape of a *single flat surface* in an object can be defined in the same way as a navigation environment. In later sections I will generalise the idea to multiple surfaces.

To recap one more time: I argued in Section 2.4 that to represent the shape of the local environment, the observer uses two related media. One holds a holistic representation of the structure of the environment's boundaries, defined both in observer-centred coordinates (the retinotopic boundary structure) and in allocentric coordinates (the allocentric boundary structure). The other holds an allocentric map of *places* and *orientations* in the current environment, that can represent environments with different topologies and/or topographies. The observer first computes an observer-centred boundary structure, and uses this to generate an allocentric boundary structure, which in turn is used to derive a spatial structure to impose on the map of places and orientations. Finally, using the observer-centred boundary structure a second time, he computes his own place and ori-

[13] Sometimes navigation is constrained by a third surface—for instance at the tip of a cube (explored either from the inside or outside). This will be a point at which the line has an *end*, either because navigation along the edge is blocked by another barrier, or because contact with the edge is lost.

entation within this newly-structured map. A very similar computation can be used to represent the shape of a flat surface navigated by the hand. But while the computations are similar, the neural medium in which they occur is likely to be very different: spatial representations of manipulable objects are well known to be held in the brain areas that plan and execute hand/arm actions, rather than in hippocampal and parahippocampal regions. I will begin in Section 2.9.5.1 by motivating the idea that parietal cortex is the place where the relevant computations take place; then in Sections 2.9.5.2–2.9.5.9 I will introduce a model of the computations themselves.

2.9.5.1 Parietal cortex as the locus of object shape representations

For navigable environments, the allocentric map of places and orientations is stored in the hippocampus, in the system of place cells. However, in the case of manipulable objects, there is good evidence that the relevant allocentric maps of places and orientations are stored in the parietal/premotor/somatosensory pathway that controls arm, wrist and hand actions—and that parietal cortex plays an especially important role. The main evidence for this comes from a condition called **object-centred neglect**, in which patients with parietal lesions systematically ignore one half of the objects they encounter (see Scholl, 2001; Chen, 2012 for reviews). Neglect can be defined in different frames of reference, often simultaneously (see e.g. Berhmann and Tipper, 1999). Object-centred neglect is defined in a frame of reference centred on the object being perceived, rather than on the observer or the environment. Typically, neglect is expressed in relation to one of the object's major axes (see e.g. Driver *et al.*, 1994). The key behavioural test for this type of neglect is that the area of an object which is ignored remains the same if the object is rotated in relation to the observer. In monkeys, there is also evidence that parietal cortex plays a central role in computing object-centred representations. For instance Chafee *et al.* (2007) gave monkeys a task in which a block had to be moved into a shape to complete it; they found neurons in parietal cortex (area 7a) which were sensitive to the block's position in relation to the object, even when the retinal location of the block varied from trial to trial. These findings argue for a medium in the brain, strongly recruiting parietal cortex, that represents a map of 'locations within a given object', that is somehow specified in a coordinate system centred on that object. In object-centred neglect, this map is systematically damaged, so that only half the locations in the object are represented.

The fact that the map of locations is implemented in the parietal/premotor cortex argues strongly that the map of places within a single (flat) surface of an object is defined in motor coordinates, as a map of possible positions or configurations of the arm (or perhaps of the body more generally) that achieve particular positions and orientations of the hand on a contact surface. But this map is given in a space that is removed from actual motor configurations, in three important ways. Firstly, the space should be configurable so it can represent surfaces with different shapes at different times: at one time it should be able to represent places (and orientations) in a square-shaped surface; at another time it should be able to represent places and orientations in a triangle-shaped surface, and so on. Secondly, the space should be able to represent shapes of different sizes. Two squares of different sizes are both squares, but the motor actions needed to interact with them haptically are very different; so there should be a size component of structure imposed on the space as well as a shape component. Finally, the space should be able to represent the surface at a range of different orientations (and distances) from the observer.

2.9.5.2 Hand trajectories for representing a coordinate system

An idea that may be specific to the representation of haptically explored surfaces is that specifications of **trajectories** of the hand provide an elegant means of defining a coordinate system. In the reach-to-grasp system in general, trajectories play a very important role. The arm has to reach a specified goal location, but it often has to reach this location by travelling along a particular route—for instance, so that the hand approaches a target object in the direction of the open palm. There are various suggestions about how trajectories are defined in the motor system. For instance, various intermediate way-points can be speicifically defined in relation to the vector between the start and goal motor states, and the hand can be constrained to pass through these points (see e.g. Oztop *et al.*, 2004), or the goal motor state can itself be perturbed, so as to draw the hand off in different directions on its route to the goal (see e.g. Lee-Hand and Knott, 2013). I will remain agnostic as to how the form of a trajectory is determined, but I will assume that it is specified separately from the goal state, as a 'deviation' from the vector linking start and goal states: in other words, a component of movement that is orthogonal to this vector.

My main point for now is that trajectories provide a natural way of representing a complete map of locations within a surface. For one thing, trajectories can be represented in a way that allows scaling to different sizes. To represent a particular shape, we need a set of trajectories with particular directions in relation to one another; the lengths of these trajectories need to have the right relationships to one another, but their absolute lengths are not important. (Note that for trajectories that are curved, the magnitude of the deviation that defines the shape of the curve can be given in relation to the length of the vector between the start state and the goal state, so this too is scalable.) For another thing, trajectories are executable motor instructions: if the observer represents a point in space as a trajectory of the hand, then he automatically knows how to make his hand *reach* this point in space. Perhaps most importantly, when the observer's hand is haptically exploring any given boundary of its navigation surface, *it is also following a trajectory.* The shape of this trajectory can be *independently defined* within the motor system, in terms of a start motor state, a goal motor state, and if necessary a deviation.

Note that the use of trajectories to define a coordinate system means that the map of places in the hand/arm is certainly structured topographically rather than topologically, although the representation of relationships *between* surfaces of a single object will be partly topological (I think).

In the remainder of this section I will describe how the same basic computations used in the hippocampus to represent a map of places in the current environment can be used in parietal cortex, within the motor system that represent trajectories, to encode a map of places in a surface currently being explored by the hand.

2.9.5.3 The hand-centred boundary structure

The first step in computing a representation of the shape of the surface navigated by the hand is to represent local geometrical features of the surface in a reference frame centred on the observer's body. In this case, clearly, the body-centred reference frame is the coordinate system of the observer's hand.

Since the hand directly contacts the surfaces whose geometry is being represented, it can represent some *local* aspects of the geometry of the navigation surface quite directly. For one thing, it can directly represent the angles between straight boundaries. When it encounters a discontinuity between two adjacent boundaries, the angle between them is given directly by the angle between the hand surfaces in contact with the two boundaries, or in motor terms, by the angle the hand has to turn in order to begin navigating along the newly encountered boundary. (This computation is exactly equivalent to the haptic computation of the angles of boundaries in a navigation surface discussed in Section 2.5.6.)

The length of each boundary can also be represented quite directly in hand-centred terms. A measure of hand speed over the surface can be derived from slip (see e.g. ??,). This can be multiplied by the time taken to travel along a given boundary to derive a measure of its length.

A common way to identify the shape of a surface is to navigate systematically around its perimeter (see e.g. Lederman and Klatsky, 1993). By navigating systematically round the perimeter of the surface, the observer can directly learn all the angles between all its adjacent boundaries, and the lengths of each boundary, both within a hand-centred frame of reference. These local measurements, defined within a hand-centred reference frame must now be converted into a representation of the shape of the surface given in a coordinate system centred on the surface.

2.9.5.4 An allocentric representation of a space defined by two opposing boundaries: the trajectory-based coordinate system

I propose that the shape of a haptic surface defined by straight boundaries is represented in just the same way as the shape of a navigation environment: by a **boundary structure**, consisting of a set of boundary pair representations. These are of the same two types: adjacent boundaries are represented by specifying the angle between the two boundaries, and the relation between their lengths, and opposing boundaries are represented by specifying the relative orientation of the two boundaries, in a way I will explain below.

As an example, say we have a surface enclosed by four straight boundaries $B_1 \ldots B_4$, as illustrated in Figure 2.8a. This shape can be represented by a single pair of roughly opposing boundaries—either B_1 and B_4, or B_2 and B_3. I will use the former pair. I will assume a coordinate system centred on the surface at corner $C1$; it remains to define the two axes of this coordinate system.

In the following account I will draw heavily on the idea introduced in Section 2.9.5.2, that the boundaries of a surface can be represented by hand trajectories, and that these trajectories can be used to define the coordinate system of a haptic surface. Each boundary

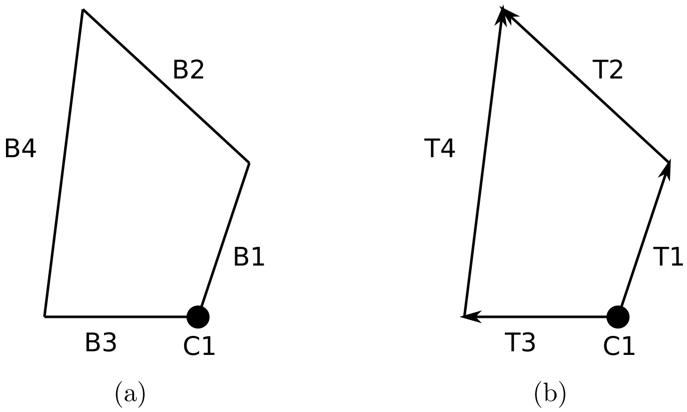


Figure 2.8: (a) A surface enclosed by four straight boundaries. (b) The boundaries of the surface represented as hand trajectories.

can be described by the trajectory described by the hand/arm when the hand moves from one end of the boundary to the other. Assume we make two separate trips from the origin corner: one along B_1 and then B_2, the other along B_3 and then B_4. The four boundaries are then described by the hand trajectories $T_1 \ldots T_4$, as shown in Figure 2.8b.

If we make two assumptions, these four trajectories provide a convenient way of referring to arbitrary points within the bounded space. One assumption is that the hand can proceed *part-way* along any given trajectory T_i. This seems realistic, since in order to move along the complete trajectory, the hand must pass through every point along it. The other assumption is that the hand can be instructed to follow a 'blend' of two trajectories: for instance, to proceed along a trajectory which combines T_i and T_j, so that its orientation, length and deviation are derived as a weighted function $blend(w, T_i, T_j)$ of the orientations, lengths and deviations of T_i and T_j. (In this expression, w is a weight varying between 0 and 1, T_i is weighted x, and T_j is weighted by $1 - x$.)

With the above assumptions, we can create two axes for the coordinate system that describes the space in Figure 2.8b: one based on weighted combinations of T_1 and T_4 (the vectors describing the opposing boundaries themselves), and one based on weighted combinations of T_2 and T_3 (the vectors linking the start and end points of the vectors that describe opposing boundaries). The axes each take values from 0 to 1 inclusive. To reach a point (x, y) in the space, we travel on a vector which is the sum of the vectors $blend(x, T_1, T_4)$ and $blend(y, T_2, T_3)$. The resulting motor coordinate system is warped to fit the surface bounded by the two opposing vectors: it can describe every point within this surface, and it is impossible to define points that fall beyond its boundaries. I will call this scheme for representing the shape of a single surface the **trajectory-based coordinate system**, and the space it defines an **opposition space**—or more precisely as an **inner opposition space**, to distinguish it from the outer opposition spaces that represent grasp affordances as discussed in Section 2.2.2.

Note that the positions within an opposition space are most accurately specified on the boundaries themselves, so the space representation scheme emphasises the boundaries of the space in the same kind of way that boundary vector cells are referred to the boundaries of a navigation environment. The intuition is that each boundary of the surface dominates the map of locations close to it.

A final assumption I will make is that an opposition space uses a measure of length that is local to the space, so that a given opposition space can describe a set of surfaces of a given shape *with different sizes*. For concreteness, I will assume that the lengths of trajectories $T_2 \ldots T_4$ are given as proportions of the length of trajectory T_1. The absolute size of the space is then given by a separate component of the representation, which I will call the **surface scaling factor**. Note that this scaling factor must still be given in a hand-based coordinate system: for instance, in units of hand-length. Thus if the surface described by Figure 2.8b has a surface scaling factor of 5, trajectory T_1 moves the hand 5 hand lengths.

The circuit that delivers an allocentric representation of an opposition space is sketched in Figure 2.9.

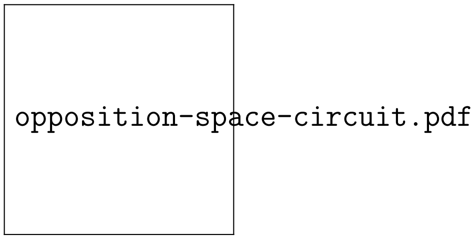


Figure 2.9: Sketched circuit for representing an opposition space

2.9.5.5 Mapping egocentric body space onto the trajectory-based coordinate system

While the trajectory-based coordinate system for a given surface is identified through hand navigation actions, it is centred on the surface: the assumption is that it stays constant if the surface changes its location and/or orientation in relation to the observer. For instance, if the observer builds a coordinate system to represent surface S when encountering S on his left, and horizontal, this same coordinate system should continue to represent S if it changes location and/or orientation.

Recall from Section 2.9.5.3 that the hand trajectories that define the boundaries of the shape are given in a hand-centred coordinate system. They do not specify hand movements *directly*, as movements of the arm: instead, they specify them quite *indirectly*, as movements of the body that cause particular movement sensations within the hand while it is navigating on the surface. The body movements that *cause* the hand to move along a particular trajectory (within its own coordinate system) will be radically different depending on how the surface is located and oriented in relation to the observer. If the surface is horizontal, the body movement that causes the hand to travel ‘forwards’ (in its own coordinate system) will move the hand horizontally away from the body. (If the observer's arm is fully extended, then moving the hand 'forwards' might actually involve *walking* forwards: locomotion actions of the whole body that move the hand in its own coordinate system are certainly be part of the repertoire to be learned.) If the horizontal surface is now rotated vertically to face the observer, the body movement that achieves the same 'forwards' movement of the hand will raise the hand upwards within its current plane. If the surface is now rotated clockwise in a plane perpendicular with the observer, the body movement that navigates the hand 'forwards' will shift the hand to the observer's right. The observer can only use a hand-centred coordinate system to create a coordinate system centred on an external surface if he has learned the complex function that maps between his own reference frame and that of his hand when it is navigating a surface. I assumed that the observer learns this mapping when he is learning about the fundamental properties of surfaces; see Section 2.3.5.1.1, and especially Section 2.3.5.1.2. If he knows this mapping for all possible surface orientations, then he can use this learned mapping to determine how to move his hand around a surface, and keep track of its location within the surface, no matter what his own spatial relationship to this surface is.

There is an important qualification to the above idea. If any time has passed since the observer originally explored a given surface and created a trajectory-based representation of its shape, it may be that he has to re-establish the location and orientation of his hand within the surface. This can be done visually, as will be discussed in Section 2.10, but within the haptic system, what is required is periodic actions that navigate along the surface's reference boundaries, that serve to reset the hand's location and orientation in relation to the surface's own coordinate system.

An interesting case to consider is when the surface moves *while* the observer is exploring it. In the worst case, this scenario would require regular actions to re-establish the boundaries of the surface. However, if he can keep his hand on the contact surface while the external movement takes place, and rely only on slip to compute the direction and magnitude of its movement on the surface, then provided he can map from motor to hand-based space, as discussed above, this movement of the surface should make no difference to his representation of the surface and his hand's location on it.

2.9.5.6 Multiple opposition spaces for representation of surfaces with complex shapes

Not all surfaces have a shape that's defined by a single pair of opposing boundaries. For instance, consider the surface shown in Figure 2.10a: there are several possible pairs of roughly opposing boundaries in this surface, but none of them represents the whole surface. Two are identified in Figures 2.10b and 2.10c. We have already seen in Section 2.5.2.2 that animals in a complex navigation space appear not to represent the complete geometry of a complex space all at once, but appear instead to attend to different aspects of this geometry at different times. I propose that a similar scheme is used to represent hand-centred spaces: specifically, that the observer represents the shape of a complex surface using a set of overlapping pairs of opposition spaces. For instance, the two opposition

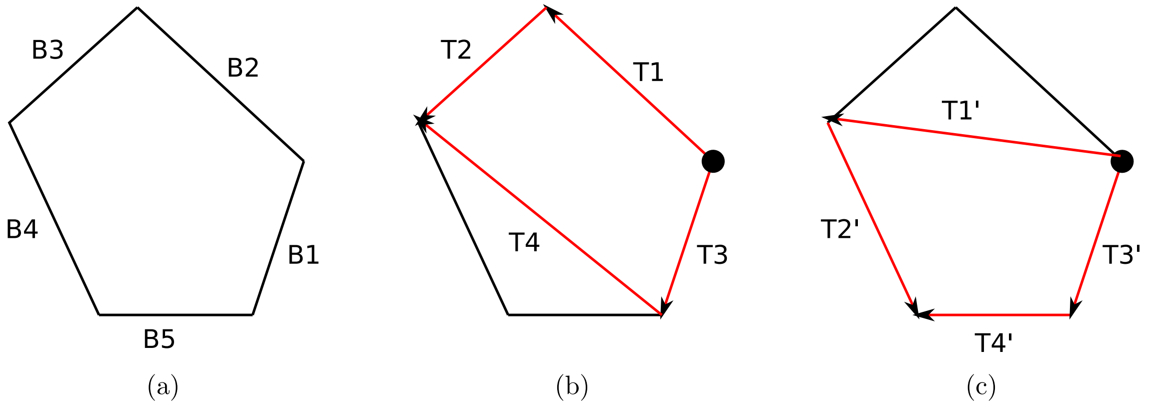


Figure 2.10: (a) A surface with a complex shape. (b) and (c) Two overlapping subspaces within the surface (shown in red) formed by opposing straight boundaries.

spaces identified in Figures 2.10b and 2.10c suffice to represent the whole shape of the surface.[14]

In the above scheme, it is important to represent not just the component opposition spaces that make up the shape, but also the spatial relationships between these spaces. If we think of these in motor terms, the important thing is to know how to *transition* from one opposition space to another. The fact that overlapping opposition spaces have shared boundaries makes defining transition relations fairly simple. For instance, to transition from the space shown in Figure 2.10b to that shown in Figure 2.10c, we need to know two things: one is the angle between T_1 and T_1'; the other is the change in the scaling factor that needs to be applied when transitioning from the first opposition space to the second. (For instance, if T_1 is 5 hand-lengths and T_1' is 6 hand-lengths, the scaling factor change is 6/5.) It is important that the transition is only made if the current position of the hand is within the region covered by both spaces. In this case, when the transition is made, the current position and orientation of the hand will be 'remapped' to a new location in the coordinate system, but this location will of course correspond to the same location within the complete surface.

The circuit that delivers an allocentric representation of a surface with a complex shape is sketched in Figure 2.11.

2.9.5.7 Curved boundaries

Using the trajectory-based coordinate system, it is possible to define maps of surfaces that have curved boundaries as well as straight ones. A hand trajectory can describe a curved line as well as a straight one. In fact I have already introduced a device which notionally

[14] Note that each individual opposition space is a little more difficult to construct by navigation, since one of the component trajectories cuts across the surface rather than following a boundary. However, boundary following routines can identify the start and end point of this trajectory.

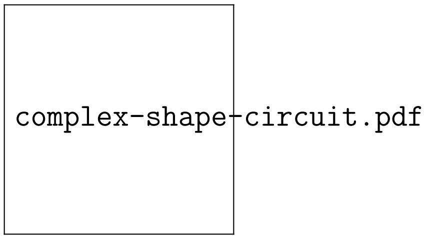


Figure 2.11: Sketched circuit for representing a surface with a complex shape

allows this, namely a 'deviation' (see Section 2.9.5.2). As mentioned there, I will not discuss the way trajectories of different forms are represented, but I will assume that it allows the definition of trajectories with roughly constant curvature. If the trajectories that define a pair of opposing boundaries are curved—within a single flat plane—then the coordinate system describes the topography of a flat space with curved boundaries.

Note that we can envisage curvature both in the trajectories that define the opposing boundaries and in the trajectories that link the ends of these boundaries. If the right (convex) curvature is specified for all four trajectories, we can define a space with a single continuously curving (i.e. circular) boundary. I will discuss related types of curved surface in Section 2.9.5.9.

2.9.5.8 Simple curved surfaces

My account of the trajectory-based coordinate system so far has assumed that the hand is navigating along a flat surface: that is, the trajectories defining the shape of the surface have all been in the same plane. But as already discussed in Sections 2.3.3.1 and 2.3.5.1.2, the hand can also travel along curved surfaces of various kinds. The curved surfaces described in those earlier sections are those with uniform curvature in two dimensions. In this section I will consider how such surfaces, and some slightly more complex ones, can be represented within the trajectory-based coordinate system.

To represent a curved surface, we just need to define a set of trajectories that do not all fall within a single flat plane. These trajectories do not need to be curved themselves. For instance, if the opposing boundaries are straight lines in different planes, and their ends are linked by straight lines, the surface between them will be curved.[15] But I will focus on cases where the boundaries themselves are uniformly curved.

If the opposing boundaries are uniformly curved in parallel planes, with the same degree of curvature, and the trajectories in between their ends are straight, we define a surface on a cylinder.[16] If the trajectories linking the ends of the boundaries are also uniformly curved, we can define a range of surfaces curving in two dimensions, for instance the inner or outer surfaces of a sphere or a torus.

[15] I think it will have a uniform degree of curvature, but I'm not sure.

[16] We can't define an actual cylinder yet, since the boundaries identify the *limits* of the surface, as well as its curvature. But see Section 2.9.5.9 for an account of closed surfaces.

Recall from Section 2.3.3.1 that surfaces with constant curvature can also be represented in a quite different way, as curvatures within hand surfaces. There is clearly a relationship between the definitions of curved surfaces given as hand curvatures and those given within the trajectory-based coordinate system: the latter scheme describes a topography of points within a curved surface with well-defined boundaries, while the former scheme describes how the hand should be shaped while traversing this surface. However, the trajectory-based coordinate system also provides an interesting framework for representing surfaces with simple kinds of non-uniform curvature. Consider, for instance, a surface defined by two boundaries, in parallel planes, with *different* (but still constant) degrees of curvature. (Again, assume the trajectories linking the ends of the boundaries are straight.) The hand curvature required to navigate at one end of this surface is much tighter than that required at the other; in fact, given that the hand occupies a region of space on the surface, the curvature of the hand must actually vary *within* its contact surface at any given time. (A natural hand pose for haptic exploration of this type of surface is one where the fingers curl in a plane parallel to the plane of curvature of the boundaries, and individual fingers have different degrees of curl.) The interesting idea here is that we can define the trajectories that link the ends of the curved boundaries as trajectories *within the hand* as well as trajectories that navigate the hand through space. The change in hand shape that must be gradually made when moving between regions of different curvature can be defined in its own right, as a hand *gesture*: that is, a trajectory of hand states with its own start and end states, and well-defined intermediate states. For instance, while the arm is executing a gesture that moves the hand towards a region of greater or lesser curvature, the hand should be executing a simultaneous gesture in which the fingers are gradually curled or uncurled, beginning with the leading finger.[17]

2.9.5.9 Closed surfaces

In topological terms, a two-dimensional surface can be 'open' or 'closed' in each of its dimensions. We have so far been considering the representation of surfaces that are open in both dimensions. But we have already mentioned examples of closed surfaces: for instance, cylinders are closed in one dimension, while spheres or tori are closed in both dimensions. In this section I will consider how closed surfaces can be represented within the trajectory-based coordinate system introduced above.

My basic proposal is that a surface that is closed in a single dimension can be represented as a special case of a space enclosed by a pair of opposing boundaries, in which (a) the two boundaries coincide in space, and (b) the trajectories that link the ends of the boundaries describe closed paths in space—i.e. paths that start from a given point and return to that same point. A cylinder is an example. (. . .) Note that we can still define an

[17] The specific gesture needed, as well as the specific hand shapes that must be assumed, depend on the orientation of the hand in the curved environment. I still need to think about how hand shape needs to change when hand orientation changes. The fact that there are constraints on how hand surfaces can be made to curve in independent dimensions is relevant: I think one consequence of this is that there are certain *preferred orientations* for the hand when it's exploring surfaces with complex curvatures.

origin within this scheme, which is useful if there are particular landmarks on the surface of the cylinder which relate in particular ways to a wider geometrical figure. But recall from Section 2.3.3.1 that there is another quite different way of representing the curved surfce of a cylinder as a hand state: specifically, the hand state that supports navigation in two dimensions, 'around' the cylinder and 'along' it. Navigation is only bounded in the latter dimension. Another way of representing a cylinder is to specify a degree of hand curvature and a *single* hand trajectory, whose start and goal states indicate the ends of the cylinder. The shapes of the planes at the ends of the cylinder are fully specified by its curvature. Using this scheme, there is no notion of an origin in the dimension of curvature, which can be useful. For instance, in this scheme there is no need to define an infinity of possible opposition boundaries: we can just define a single opposition boundary within the plane of the hand.

I suggest that a surface that is closed in both dimensions can also be represented as a space enclosed by opposing boundaries, if we define the opposing boundaries as each having a path that curves outwards, and the trajectories that link the ends of the boundaries as being null trajectories, whose goal states are the same as their start state. This means that the opposing boundaries touch each other at each end, but are separated in the middle. This projection of a sphere onto a two-dimensional plane over-represents locations at the 'poles', but there are many spherical shapes in nature where it that have radial symmetry about one axis, where this kind of over-representation is appropriate: for instance, the segments of an orange are smaller at its two ends than around its middle. The projection is also convenient in that an origin can be defined. Again, there is an alternative definition of a sphere that provides no origin, and does not cause any distortion of the surface: that is to define it as a hand surface with a constant curvature in both dimensions. Different hand shapes specify different degrees of curvature, and thus different sizes of sphere (in hand-based coordinates). Again, this way of representing a sphere is convenient for representing opposition spaces, because they can be defined directly in hand-centred coordinates, and applied to a sphere no matter how it is oriented.

2.9.6 Representing the geometry of pairs of surfaces

Now that we can represent the shape of single surfaces, we need a way of representing the spatial relationships that link these surface representations together. I will first consider the relation between a pair of adjoining surfaces, and then consider the relation between a pair of opposing surfaces.

2.9.6.1 Adjoining surfaces

Recall from Section **??** that the observer can activate different spatial representations of the current navigation surface at different times, that represent different sub-spaces within this surface. The operation that implemented a switch between different representations was termed a *transition action*. A key idea was that a transition action involves 'pivoting' around one of the boundary representations active in the current spatial representation, that is also involved in the new spatial representation, though it plays a different role. In this section I will propose that the same notion of a transition action, marginally extended, can serve to represent the relationship between two different surfaces in an object, provided they are adjacent (i.e. that they share a boundary). My proposal is that the agent represents this relationship in hand-centred motor coordinates, in relation to the haptic action that moves the hand from one surface to the other.

In the simplest case, two adjoining surfaces meet at a discontinuous edge. I will first consider this case. Then I will consider the case where the surfaces merge continuously into one another.

2.9.6.1.1 Surfaces meeting at an edge As discussed in Section 2.9.4, the hand can encounter discrete boundaries in a flat surface, which can be explored haptically. If the boundary is concave, encountering the boundary directly activates an additional surface in the hand; if the boundary is convex, encountering the boundary allows additional surfaces to be activated by adjusting the pose of the hand. In either case, the boundary provides an interesting opportunity to move the whole hand into the newly encountered surface. Say the hand is currently exploring external surface S_1 using hand surface s_1, within which there is a hand-based coordinate system $C(s_1)$. Say the hand now encounters external surface S_2 using hand surface s_2. I will define a **reconfiguration action** as an action that repositions the hand (using the arm, and maybe other motor systems) so that hand surface s_1 achieves stable contact with external surface S_2.

A reconfiguration action is an action that is outside the normal repertoire of actions that move the hand surface within a continuous surface. It could involve a sliding action, as in the case where a hand slips from one surface of a cube to an adjoining one. It could involve a more holistic change, in which the hand is disengaged from one surface, rotated (and perhaps shifted), and repositioned in the new surface. In many of these actions, stable contact with the relevant surfaces may be temporarily suspended.[18] But while it temporarily violates the 'rules' about how to navigate s_1 within a continuous surface, it re-establishes these rules within a *new* surface: after the action the hand is on a new surface, in which the hand-centred directions 'forward', 'back', 'left', 'right' etc are defined. In objective terms, a reconfiguration action rotates the hand's reference frame by a certain angle, about an axis parallel to the boundary between the two surfaces.

Of course the observer has to learn what body actions achieve a given transition action. He has to learn what the various different angles defined within the surface of the articulated hand mean in motor terms. Again, this is something that depends in a highly nonlinear way on the position and orientation of the connected surfaces, as well as on their angle. But the important thing is the angle that the hand goes through, defined in a hand-centred frame of reference.

When the observer's hand arrives in the new surface, he will have to find a way of representing the shape of this surface, which will again involve exploration of the new

[18] Are there similar actions in navigating around one's local environment, when the observer shifts attention from one surface to another? A momentary loss of balance?

surface's boundaries. This shape has to be related to the shape of the previous surface. If we imagine flattening the two surfaces, so they form a single continuous surface, we could describe this relationship by a transition action pivoting around the boundary between the two surfaces, that orients the coordinate system of an opposition space in the first surface to that of an opposition space in the second surface. I propose that a pair of adjacent surfaces in an object, connected by a boundary, is defined as a transition action that links an opposition space on the first surface to one on the second surface—*plus* a specification of the angle of rotation of the hand as it pivots on the boundary from one surface to the other. This transition does not just involve an update to the currently active spatial representation, but also an actual action of the hand.

The observer's hand can now navigate over all the surfaces on the object, and build a representation of its complete 3D shape. Of course when navigating, the observer is likely to re-encounter places he has been before: that is, to re-establish opposition spaces that have been established previously. I assume that the observer can recognise when this happens. There are two methods which allow this. One is to match the opposition space activated through current haptic exploration against stored representations of opposition spaces encountered elsewhere on the object, using the same mechanisms that support recognition of the current environment (see Section 2.16.1.1). The other assumes that the external surface does not move during navigation: in this case, places on the object can be associated temporarily with body-centred motor states (e.g. positions of the arm), and re-encountering a given body-centred position can directly indicate reactivation of a given opposition space. No matter how opposition spaces are reactivated, the above scheme means that the representations of transitions between opposition spaces form a *graph*, rather than a simple linear sequence, in which there can be several different ways of navigating from one opposition space in the object to another. I will talk more about this idea in Section 2.9.8.

2.9.6.1.2 Surfaces connected by a continuous boundary Recall from Section 2.3.3.1 that the haptic modality can generate simple and useful representations of curved surfaces, if these have constant curvature, in one or both dimensions. A complex 3D surface may, however, have different degrees of curvature at different points. We have already considered the case where curvature varies monotonically as a function of position within the surface: as discussed in Section 2.9.5.8, this can be well modelled within the trajectory-based coordinate system. But there are other cases where a curved surface is better modelled as having discrete regions with different constant curvature, connected by an intermediate region. In this case, we will naturally represent the two regions as separate regions, with their own associated coordinate systems. The boundaries of each region will be defined at the point where stable contact is lost, i.e. where the hand needs to change its shape in order to maintain contact with the surface. Of course, the intermediate region between the boundaries of the two sub-surfaces also needs to be represented.

Here, I suggest we need to define a special class of reconfiguration action that describes the intermediate region. Exactly what this is will be a complex function of the two curved surfaces and their spatial relationship—I assume the observer has to learn the function for each pair of surfaces. My hope is that there are certain classes of surface pair that occur more regularly in nature, and that learning can focus on these: for instance, the surface of the body that links the shoulders and the neck (which each have roughly constant curvature), or the surface formed by a junction between two branches of a tree. Remember that we only need to learn *approximate* representations of these intermediate regions, that suffice to support navigation along the surface from one region of constant curvature to the other. The physical surface itself helps to establish a precise hand curvature, since the hand is not rigid, and is being pressed against the surface.

2.9.6.2 Opposing surfaces

As already discussed in Section 2.2.2, there is evidence that the observer's affordance-based representations of 3D object shape also make explicit the various opposing surfaces in the object that support stable grasps. In Section 2.2.2 these pairs of surfaces were termed 'outer opposition spaces', following Iberall and Arbib (1990). They are like the inner opposition spaces defined in Section **??** in that they involve pairs of roughly opposing boundaries; however, outer opposition spaces represent these boundaries in 3D rather than 2D, and represent them 'from the outside', from the perpective of a motor system that has two opposing contact surfaces of its own, that can be somewhat independently positioned and oriented in relation to one another.

The foundational idea is that the observer can explore an object somewhat separately with two separate body surfaces in the haptic system simultaneously. These can both be on the same hand, or they can be on different hands, which both contact the same object. During this type of exploratation, a special situation can arise in which the hand surfaces are navigating within roughly opposing boundaries on a single object. In this case, applying 'downward' force within one surface (in its coordinate system) will be registered as 'upward' movement within the other surface (again in its own coordinate system). This indicates that the two surfaces are part of a single rigid body. I propose that this state is recognised by the activation of a special signal, the opposing-pressure signal. In a particular sub-case of this case, the hand surfaces achieve stable grasp on the object: this is when enough pressure is generated on the opposing surfaces that it can be moved, or lifted. This has its own special signal whose purpose in learning reach-to-grasp actions I have discussed in detail elsewhere (see e.g. Knott, 2012; Lee-Hand *et al.*, 2013). For present purposes, the main issue is to specify how the representations of 'outer' opposition spaces associated with these signals can be integrated within the topological structure of adjacent 'inner' opposition spaces discussed in Section 2.9.6.1. The tricky thing is that the spatial encoding system I have described can only represent one surface at a time.

My proposal is that the agent navigates round the surface using a single 'dominant' or 'reference' motor surface, and that the spatial representation normally encodes the object surface contacted by this motor surface. A secondary motor surface can also travel around the object, guided only by local haptic cues of the kind described in Section 2.3. If this secondary exploration ever results in activation of the opposing-pressure or stable-grap signals, I assume that the current contact surface of the object is temporarily stored, and the surface contacted by the secondary motor system is identified. Then the two surfaces can be linked via a special type of transition action associated with opposition surfaces, so that when the agent has contact with one surface, the topological structure of inner opposition spaces indicates how he should orient his hand to establish an outer opposition space with another motor surface.

2.9.7 Representing the location of external objects within a surface

When exploring a given surface of an object, the observer may encounter another object on it. For instance, when he is exploring a tabletop with his hand, he may encounter other objects on this surface. He must be able to store the location (and orientation) of these other objects on the surface, in the same coordinate system he uses to represent the location of his hand wihtin it. However, there must be a clear way for him to *distinguish* between representations of his own hand's location on the surface and representations of the locations of other objects on the same surface.

This issue is exactly analogous to the issue of how the observer distinguishes between his own location in his navigation environment and that of external individuals in this same environment, as discussed in Section 2.6.1. In the case of environment-based representations, my proposal in that section was to envisage two media for representing the 'current place': one that is hardwired to represent the observer's place, the other that represents the place of a currently attended 'subject', which can in one mode be the observer himself, and in another mode be an external individual in the same environment as the observer. In order to identify the location of an external individual, the observer needs to refer to his own location and orientation in his environment. The case is no different in the haptic system—only here, since the external individual is identified directly by making contact with it, the location of this individual relates directly to that of the hand: at a first approximation, is in fact *identical* to the location of the hand.

To make the relevant distinction within the haptic system, I will again assume two cognitive modes: a self-perception mode, where the observer is monitoring his own hand as a moving object within the surface being explored, and an external perception mode, where the observer is using his hand as a sensory instrument to identify the location (and shape) of external objects occupying the same surface. Envisaging these two modes is a helpful way of confronting the duality of representations generated by the somatosensory cortex. As discussed in Section 2.3.3, somatosensory cortex represents surfaces within the observer's body, but also surfaces within external objects. Any given tactile stimulus can inform the observer both about himself and about an external object: keeping these two kinds of representation separate is very important.

In self-perception mode, the subject is the observer's own hand: the observer can represent the location of his hand within its own navigation environment, or complete navigation actions (as will be discussed in Section 5.2, and possibly Section 5.4.2). In external perception mode, the subject is something *contacting* the observer's hand. I assume that when exploring an object, the observer is in self-perception mode by default, and that a special kind of sensory stimulus that I will call 'object-encountering touch' triggers a transition to external perception mode. I will first discuss characteristics of the stimulus, and then discuss the transition it triggers.

2.9.7.1 Object-encountering touch

An object-encountering touch is a touch on some surface of the hand other than the one contacting its navigation surface—for instance on the fingertips. In this respect it is similar to the sensation of contact with a 'barrier' delimiting an edge of the navigation surface (see Section 2.9.4.1) An object-encountering touch is defined as a touch sensation on a non-navigating surface of the hand that occurs *away from* the boundaries of the surface being navigated.

Sometimes there are other characteristics of an object-encountering touch. If the object is made of a different material than the navigation surface, the touch may be associated with a characteristic sound. (For instance, the sound made by the hand contacting a cup or a glass is different from the sound it makes when it travels over a tabletop.) If the object is moveable, then the initial touch may cause it to move: this movement could be picked up haptically, as transitory stable contact sustained during movement of the hand, or possibly through audition, since objects sliding on a surface often make distinctive sounds.

2.9.7.2 Establishing a salient location by touch

However the object-encountering touch is registered, I propose that its effect is to transition the observer into external perception mode. In practice, the main effect of this is that the current location of the hand is copied into a new medium holding a map of salient locations in the current surface. While a map of salient locations is created in parallel by the visual system, it is created more serially within the haptic system. However, since there are many surfaces within the hand, it can potentially encounter more than one object at a time—and given that two hands can explore the same surface somewhat simultaneously, it is possible that multiple salient locations can be identifed at the same time. Nonetheless, I envisage that once a salient location has been identified, the observer can continue exploring the surface seeking additional salient locations if he chooses, establishing something quite like the saliency map produced by the visual system. At any point, of course, the observer can also choose the most salient or interesting of these locations for further exploration. The key point is that when the observer first touches an object, he knows there is something there, but not what it is: serial attention must be deployed to this object in order to determine this.

2.9.7.3 Deploying serial attention to a salient location in the haptic system

To deploy serial attention to a salient location, a winning salient place must be picked from the map of salient places. In the haptic system, the effect of this is that the observer's hand stops moving within its current surface and starts a haptic exploration of the external object within this surface at the selected place.

The transition needed to accomplish this is in one sense similar to that needed to navigate from one surface of the current object being explored to another (see Section 2.9.6.1). However, a tangible surface encountered within a surface rather than at its boundary is represented as an object in its own right, indexed to the surface, rather than as a feature of the surface itself. For example, the *rim* of a cup is a boundary of one of the cup's surfaces, that partly defines this surface, while the *handle* of a cup is an object within one of the cup's surfaces, occupying a particular position and configuration in relation to it. Whether the object is fixed to the surface, as in the case of a handle on a cup, or moveable, as in the case of a cup on a tabletop, the object is still indexed to the surface at a particular place and orientation, resulting in a spatial relationship between two complete object representations. These will be considered further in Section 2.15.

2.9.8 'Complete' haptic representations of object shape

I have now discussed many components of haptic representations of the shape of manipulable objects. In this section I will step back and consider haptic shape representations as wholes: what sort of thing is a haptic representation of a complete object?

2.9.8.1 A definition of the allocentric boundary structure that represents a 3D shape

The main idea here is that a 3D shape can be described as an allocentric boundary structure, made up of pairs of opposing boundary surfaces for a given surface, plus pairs of angles between adjacent surfaces that allow transitions from one surface to another.

2.9.8.2 The deictic nature of 3D shape representations

2.9.8.3 The topological structure of 3D shape representations

An important idea is that the above definition of an allocentric boundary structure is at the higher level a essentially a *topological* one. The angles between surfaces define a graph of possible haptic transitions between surfaces. This graph may represent

Obviously, I should refer back to Section 2.8.2 here.

In relation to the idea that an object shape is represented by sequences of precompiled transitions between adjacent surfaces, note that damage to the cerebellum is known to impair representation of shapes in the haptic modality; see e.g. Shimansky *et al.* (1997).

2.9.8.4 Implicit representations within a 3D shape representation

Here mention the idea that although the map medium represents one opposition surface at a time, the full 3D geometry of an obect is represented *implicitly* by an allocentric boundary structure. The representation supports operations that transition between representations of related surfaces. When we think of a cube, do we think of all aspects of its geometry at once? My suggestion is that we only think of specified aspects, but are able to transition very rapidly

This is a point that relates to one made by Damasio in relation to memory in general. Damasio sees LTM structures as encoding *potentials* for representations rather than representations themselves. A useful way of thinking about an allocentric boundary structure is as a related set of *potential* spatial representations in the parietal motor modality, rather than as a single, holistic spatial representation in this modality. Even vision, with its parallel access to distant locations and its ability to represent complex forms, can only ever directly represent a *portion* of the complete geometry of a given object. And my suggestion is that visual attention often focusses on particular aspects of the geometrical representations computed by vision.

A very similar point can be made about our apprehension of our local environment and its contents. We have the impression that we access information about the contents of the current 'scene' in parallel, but it is well known that this is an illusion (see e.g. Henderson and Ferreira, 2004). My proposal, as mentioned at the outset, is that representations of scenes and objects use the same basic scheme, and that there are many parallels between them: the idea that information about the geometry of a single object is accessed through rapid attentional transitions to different parts or aspects of the object is a case in point.

2.9.8.5 Redundancy in 3D shape representations

There is a lot of redundancy in the representation of 3 shapes in the scheme just outlined. There is redundancy in the way single surfaces are represented, in the sense that different opposing boundary pairs can often overlap extensively. There is also redundancy in the way spatial relations between surfaces are represented, because there are typically several different ways of navigating from one surface to another. But this redundancy is not harmful—in fact it is positively useful, because the spatial representations can directly inform strategies for manipulating objects, favoring the strategies that have been most highly learned. If the hand is on a given surface and needs to reach another surface, we can look for the path in the graph that

2.9.8.6 The coordinate system of a whole object

We have discussed the coordinate systems of individual surfaces within an object. But it is also important to identify a coordinate system for the *whole* object. This is partly because whole objects often have their own local representations of 'up/down', 'front/back' and 'left/right'. (The latter vector is often less important, since many natural and man-made objects are roughly symmetrical in the plane defined by the 'up' and 'forward' vectors, meaning that their left-hand side is a mirror image of their right-hand side. But there are still important distinctions between left and right—for instance humans and other animate agents can move their left and right effectors independently.[19]) But even if there

[19] Within certain limits; see e.g. Liz Franz's work.

is no strong object-centred coordinate system of this kind, it is still important to be able to represent the spatial relationship of an object with its own containing environment.

I assume that the coordinate system of a complete object is simply identified with the coordinate system of one of its largest surfaces. Often there is an obvious candidate for this surface. If there are several equally good candidates, I assume one is selected more or less at random.

As discussed in Section 2.9.5.9, surfaces with rotational symmetry often only need to be defined in relation to one directional dimension, plus a second dimension indicating the local curvature, which is a constant. In cases like this, I assume that the other dimension can simply be left undefined; I will discuss how this could be implemented in Section **??**.

2.9.8.7 Relative size and shape properties

In the context of the above discussion, it's useful to think about properties referring to size (specifically 'big' and 'small') or to some combination of size and shape (such as 'tall', 'short', *long*, *wide*, 'thin' and 'fat'). These should be representable in relation to the primary axes of an object's coordinate system. Say something about this here, and then refer to Section 2.11.2.

2.10 Visual representations of 3D object shape

2.10.1 Background: neural representations of object shape

I mentioned in Section 2.3.5.1.3 that macaque V4 contains representations of surfaces with different curvatures. In fact

Refer to Amedi *et al.* (2010);

Refer to Lacey *et al.* (2009) on the idea that representations in LOC (arguably the human homologue of macaque V4) are inherently multimodal, not just visual.

Refer to Yamane *et al.* (2008) for IT computing a code for 3D object shape..

Yamane *et al.* (2008 also found that IT neurons preferentially encode particular *combinations* of local shape fragments, in line with classical geometric models of neural object representations (Marr, 1982; Biederman, 1985).

3D aspects of object shape: refer to Orban (2011) for a detailed model.

Refer to Krause *et al.* (2013) for work in machine vision using 3D object representations to represent the locations of object parts.

2.10.2 The main idea

I think the basic idea in here is that vision can also selectively deliver representations of portions of objects. Visual representations are mapped onto active somatosensory representations, and these latter representations are partial, and extremely context-dependent; we therefore expect vision to

2.10.3 Attention to contact surfaces on objects

The classical idea is that vision delivers ‘a set of motor affordances’. The point should be that these affordances are defined relative to the hand’s current position on an object, and that separate affordances can be computed in parallel, so that one particular one can be selected. This doesn’t just involve selecting an action: it also involves selecting a visual region within an object, or something like that—for instance the region where the fingers should make contact with a particular surface on an attended object.

[I like the idea that the observer creates not just one saliency map to represent places in the currently attended object, but *several* saliency maps, which compete amongst one another, that represent different surfaces within the object. The saliency maps are ultimately tied to separate effector-based motor systems, and are computed in parallel; the representation of the spatial relations between the surfaces is given by the representation of angles between the different effector-based motor systems.]

2.10.4 Parallel, pre-attentional identification of proto-objects

You can refer e.g. to Peterson and Kim (2001). More latterly, Kimchi and Peterson (2008).

Somewhere in here, also refer to the idea that infants learn to orient the major axes of objects perpendicularly to the line of sight (see e.g. Pereira *et al.*, 2010). Also mention that IT computes axis-based shape representations (see e.g. Hung *et al.*, 2012).

2.10.5 Marr’s pathway

In Marr’s (1982) classic model, vision first identifies surfaces in a scene (in a ‘2$\frac{1}{2}$d sketch’) and then combines surfaces together into

2.10.6 Visual attention to objects and surfaces

In relation to borders, refer to Zhou *et al.*’s (2001) study on border ownership in monkeys. This is an awesome study.

There’s also evidence that regions on a surface near its corners are processed preferentially (see e.g. Bertami *et al.*, 2013).

In relation to surfaces, refer classically to He and Nakayama (1995); more recently Ciaramitaro *et al.* (2011).

In relation to objects, refer to a recent review by Chen...

2.10.7 Evidence for surface-centred visual representations

Olson and Gettner is a good example of visual representations that are somewhat invariant to rotation of a perceived object.

Recently there has been some evidence of translation-invariant visual representations—or at least of the influence of such representations on low-level vision. This comes in particular from a study by O'Herron and von der Heydt (2013), showing that the border ownership properties of V1 cells activated when an object's border is in one retinotopic location are transferred to V1 cells in a different retinotopic region when the object moves on the retina.

Related to the O'Herron and von der Heydt paper, you can also refer to human data from Demeyer *et al.* (2011) that show that perceptual grouping of object contours survives saccades, again suggesting the influence of an object-centred representation of the relevant surfaces created by the grouping process.

2.10.8 Links between visual and somatosensory representations

On the mapping between visual and somatosensory representations: refer to Haggard *et al.* (2007), Bolognini *et al.* (2011). See also Gillmeister and Forster (2012).

This point relates to notions of visual attention, and especially to ideas about object-based attention. As already mentioned, neglect is expressed in relation to one of the object's major axes (see e.g. Driver *et al.*, 1994).

Also refer to the idea that location is encoded hierarchically in relation to objects (see e.g. Baylis and Driver, 1993).

2.10.9 Hand position biases on object-based attention

Cosman and Vecera (2010) showed that an agent's hand position exerts an influence on figure-ground assignment: specifically, visual regions associated with the targets of reach actions were more likely to be identified as figure than visual regions not reached towards.

2.10.10 A computational model of object/surface identification

A good model that identifies surfaces is that of Jehee *et al.* (2007).

The upshot of this algorithm is a subset of selected, active visual representations right through the visual pathway, from low to high areas. There are two components to this pattern of activation: one identifies the form of the selected object or surface; the other represents visual stimuli 'within' the selected object or surface. This latter representation is simply a portion of the retinotopic map bounded by the boundaries of the selected form.

2.10.11 Computing an object-centred visual representation

The patterns of activation in the visual pathway representing the form of a selected object/surface and the set of locations within this form provide the input to another algorithm that can map these visual representations to object-centred representations.

I think the most plausible algorithm might be to use basis functions (see e.g. Deneve and Pouget, 1992). The motor representation is explicitly object-centred: it is invariant

2.10.12 Visual representations of outer opposition spaces

Since vision delivers information about the surfaces of an object in parallel, the agent can generate a holistic visual representation of an opposition space in an attended object, and learn to map this onto a holistic representation of a pair of motor surfaces.

2.10.13 Visual routines that identify a haptically established surface

I propose that visual attention tracks the location at which the hand surface makes contact with an external surface, while it is exploring the surface and maintaining a constantly active stable support signal. This

2.10.14 Structuring retinal space to map it onto a map of places

When you attend to an object 'as an environment', I propose that there's a visual analogue of the operation that structures the map of points in space to define the topology/topography of the environment, so that points are created differently in different planes of the object, somehow. The visual attention medium is still looking for object-like things within this space, but there are operations that attend to portions of the space, like surfaces, edges, corners.

So: what are these visual operations / representations?

2.10.15 What's going on in IT?

The ventral 'what' visual pathway is the one responsible for our ability to 'recognise' visually objects, rather than interact with them (see e.g. diCarlo *et al.*, 2012). What information is computed in IT, that is not computed in the parietal/premotor pathway? Whatever it is, it certainly still involves shape to some large degree, not just a representation of 'semantic category' divorced from shape (see e.g. Baldassi *et al.*, 2013). But it's a shape representation that is much better at discriminating different types of object (see e.g. Lehky and Sereno, 2007).

I guess one thing is that IT calculates invariances over pose, while parietal/premotor cortex cares a lot about pose.[20] While IT neurons often have pose-specific firing rates when shown 2D figures, when shown 3D figures, they show a considerable degree of pose-invariance (see e.g. Hung *et al.*, 2012).

Perhaps a better idea is just that IT takes the visual shape representations that have their origins in mappings with the haptic system, and finds new ways to exploit them, for semantic and goal-based purposes. Shape is useful for manipulation, but also for other things. Or perhaps: the parietal pathway tells you how to interact with the global shape of a perceived object; but the inferotemporal pathway tells you about aspects of shape which

[20] But not when it comes to hand-based representations. Your whole idea is that haptic represenations of objects are also somewhat invariant.

are not relevant for these interactions (but would naturally become relevant for interaction with sub-components of the object, like the handles on the door of a car, or the rims around the edges of its doors).

2.11 Representations of object shape in LTM and WM

As already discussed in Section 2.7.2.1, I propose that individual objects perceived by the observer are represented as *LTM individuals*, patterns of activation in a convergence zone assembling the perceptual and motor properties of objects. An LTM individual is a sparse assembly of neurons, probably in a parahippocampal medium, which represents an individual by being linked to a range of assemblies in more sensorimotor areas of the brain that represent phenomenal aspects of this individual. One of the important aspects of a concrete individual is its shape. (Further dimensions are discussed in Section **??**.) For present purposes, the main idea is that each LTM individual representing a manipulable concrete object is linked to an assembly in the reach/grasp pathway (and in particular in parietal cortex) that represents its shape as an allocentric boundary structure.

In the remainder of this section I will consider a few ramifications of this idea.

2.11.1 Recognising an individual object

When the observer perceives an object, either haptically or visually or both, he builds a representation of its shape—probably a partial one. Somehow this partial representation can be matched to a more complete stored representation of shape in the parietal cortex, which is in turn associated with a LTM individual.

When this happens, it is essential that the observer also identify his current relation to the perceived object, i.e. its pose.

2.11.2 Size and shape representations for types of object in LTM

It must be possible to represent the shape of a *generic* cup in LTM. Relative size and shape adjectives like *big*, *small*, *tall*, *short* and so on (see Section 2.9.8.7 make reference to these measures. How do they work? Explain that in this section.

LTM individuals represent tokens individuals. I propose that alongside these is a set of **LTM types**, one for each basic-level type. Importantly, I assume that LTM types are represented in the ventral/inferotemporal pathway—but *linked* to shape representations in the parietal pathway. (This is an idea that's very much to the fore in the model of Fagg and Arbib, 1998, in line with Milner and Goodale's observations about patients with damage to parietal cortex, who can still preshape their hands for grasping a cup with *typical* size, shape, and pose.) When you recognise the type of an object, you activate this shape; you also activate the actual shape of the object. This means that you can directly evoke a representation of the differences between the shape of the current token object and that of the prototypical object of its type.

The differences can be recorded using the axis-based coordinate system of the object, as discussed in Section 2.9.8.7. (Elaborate on this idea.)

An important consequence of this mechanism for computing differences from a prototype is that the shape of a token object can be *stored in LTM* as a set of deviations from the shape associated with its type. My idea is that the most elaborate, exhaustive shape representations held in LTM are associated with LTM types, and that the shapes of most token individuals are stored as deviations from the relevant types. Of course we can still represent the fact that token individuals have particular sub-parts, contents, possessions and so on. The point is just that these parts are indexed to a shape representation that is derived from a generic shape template, rather than being represented from scratch for each token individual.

How are these differences identified? I assume there's a match operation that maps a shape representation derived perceptually onto a shape representation stored in LTM. This is a tricky operation, because shapes are represented in a recurrent circuit that only makes explicit one dimension or one surface at a time, as discussed in Section **??**. I don't have a suggestion about what this match operation looks like yet. My guess is that the match can be done directly, on the basis of similarity between high-dimensional feature vectors, but that when a match has been *found*, we can step through the recurrent shape representations of the token individual and its matching type *in parallel*, and identify the particular dimensions of variation during this process. This might explain why size adjectives are structurally higher than other shape adjectives cross-linguistically, as in the English *big fat*, **fat big* (see classically Cinque, 1994).[21]

For instance, there's an LTM type for the type 'cup'. This is associated with a shape of the same kind as

2.11.3 Haptic representations of familiar and unfamiliar objects

[This may be out of place: it should probably be earlier in the section.]

Recall from Section **??** that representations of the shape of the observer's environment in PPA become more viewpoint-independent as the observer becomes familiar with it. Something similar is true in the brain areas that create somatosensory representations of shape: haptic experience with both familiar and unfamiliar objects activates LO, but the networks that LO participates in are quite different for familiar and unfamiliar objects (see Lacey *et al.*, 2009 for a summary).

[21] This may apply within shape adjectives too. *Long fat sausage* is better than *fat long sausage*; *tall round building* is better than *round tall building* etc. On this analysis, this kind of ordering constraint has its origin in cognitive representations of 3D object shape, rather than in anything specific to language.

2.12 Representing the spatial relationship between an object and its environment

In Section 2.6.2 I sketched a function that identifies the location of a salient external object in the observer's current environment. As discussed in that section, this function must deliver not only the location of an object but also a representation of its spatial relationship with its environment, including its orientation in the environment and the surfaces in the environment that support it or constrain its movement. The model of object shape outlined in Sections 2.9 and 2.10 provides the basis for this representation.

2.13

2.14 Representing objects with hierarchical structure

2.14.1 Representing parts of an object

A cup has a handle. Even if you broke the handle off the cup, the cup would still be a cup and the handle would still be a handle, so in some sense we can consider these separate objects. I'll define that by saying that the two objects are represented by distinct LTM individuals.

A key idea now is that the spatial representation of an object discussed in Section **??** provides a map of *places within* the object (or places in its environs). This can be very straightforwardly thought of as an environment *in which other objects are located.* This is a very attractive property of a spatial representation system that works the same way for objects and environments. Every object can also be thought of as an environment.

Note that the spatial representation of an object allows us to specify not just the location of a sub-object but also its orientation. The sub-object will have its own coordinate system, just like the hand, with its own front/back, left/right and up/down axes: in the same way as we can

I'll sometimes use the term **indexing** to describe the relationship between an object and a sub-object occupying some space within

2.14.1.1 Articulated objects

Since the transitions between surfaces involve rotating the hand through specified angles, we can easily extend the scheme to describe articulated objects, whose parts can rotate in relation to one another. This allows us to represent the articulated bodies of animals, including humans: the human body is of course a special case.

Note that if a human agent has a cup within his grasp, this cup is indexed

2.14.1.2 Broken objects

If I break a cup, I create separate objects, but these can be referred to as *pieces* of the object (for instance pieces, or shards, of a broken cup). There's an interesting concept of a 'piece' that I'd like to think about here.

For instance, there's an important difference between pieces and parts. Parts are geometric features of a complete object: there's no guarantee that an object will break into its parts. (For instance, when a cup is broken, the rim of the cup may simply cease to exist.)

2.15 Spatial relationships between objects

2.16 Representations of object location in LTM

2.16.1 LTM environments

2.16.1.1 Recognition of a perceived environment

Consider the process by which an agent identifies his current environment as one he is familiar with (i.e. which is already represented in long-term memory), based purely on its shape. This involves matching his perceptual representation of the environment's boundaries with an allocentric boundary structure stored in LTM. If we assume the agent is disoriented, he has no prior expectations about how his retinotopic boundary representations map to allocentric boundary structures: this means that for any allocentric boundary structure there may be several possible 'fits': a square environment is ambiguous 4 ways; an oblong one is ambiguous 2 ways, and so on. Each fit is consistent with a specific representation of the agent's location and orientation in the hippocampal system. If there are cues other than shape (for instance, a distinctive landmark on one boundary), these can serve to reduce or eliminate ambiguity. Noticing such cues belatedly can result in a sudden, discrete reorientation, implemented as a sudden change within the system of place cells and head direction cells in the hippocampal system.

2.16.2 LTM individuals

2.16.3 Object location memory

2.17 Old below

2.18 Transitions between environments in experience and LTM

[This section will introduce the topic of transitions. Subsequent sections will describe particular aspects (e.g. to do with learning etc).]

The preceding sections provide all the preliminaries for describing the complex topic of environment transitions, including:

- LTM update operations
- Locomotion actions
- The base motor system
- The support affordances function

2.18.1 Moving into a new environment

2.18.2 Attentionally entering a new environment

This is what happens when you decide to establish an attended object as an environment. In this operation, the current environment is initially the environment that the attended object is in. This environment defines a topographic map of places, and the attended object is represented as being 'at' one of these places. When the observer attends to this object, he represents it as a collection of surfaces: something analogous to a set of boundaries, but perceived 'from the outside'. However, the observer can choose to execute an operation that establishes the attended object *as the current environment*. As usual, this operation has effects on perceptual representations and on LTM representations within the environment/place system. The key effect within the perceptual representations is that the topographical map of places starts to represent places *within the attended object*, rather than places within the agent's actual environment. This allows us to represent the location of things 'in', or 'on' the object, or which stand in other relationships of stable support or containment with the object. The key effect within the LTM system is that a new LTM environment is selected as the current LTM environment. Before the operation, the current LTM environment is the one the observer is in; afterwards, the current LTM environment is a new unit that represents the attended object *as an environment*. I'll introduce this idea in more detail, and discuss some of its interesting consequences, in Section **??**.

2.18.3 Jumping to a new environment

A final way of transitioning to a new environment involves establishing a distant environment as the current environment, relying purely on his LTM representations. This may happen when the observer is thinking about a distant place at the current time, or having a memory of occurrences at a distant place in the past, or making plans about a distant place in the future, or simply imagining a distant place at an unspecified time. In all these scenarios, the media that normally hold a perceptual representation of his environment (specifically, a representation of the current environment's boundaries, and the map of places which these boundaries collectively define) are freed up to represent the structure of the distant environment. the element

2.19 Representing objects as environments

Now that I'm allowing objects to be represented as environments, it's useful to distinguish between 'normal' environments like gardens, rooms and streets, and 'objects-reconstrued-as-environments'. I will call environments like gardens, rooms and streets, **gravity environments**, or more informally, **locomotion environments**. They are environments which are predominantly horizontal surfaces, in which agents support themselves themselves by positioning their centre of gravity appropriately in relation to a base motor system. The three-dimensional components of these environments are for the most part vertical obstacles or barriers, which 'contain' by constraining his horizontal movements. I will call objects-construed-as-environments **convex environments**. These environments are ones which the agent is on the outside of, rather than on the inside of: the agent interacts for the most part with their external surfaces rather than their internal ones. The distinction between gravity environments and convex environments is only a rough-and-ready one, but it will be helpful in structuring the discussion.

2.19.1 LTM individuals and LTM environments

This is where I say that there's a 1:1 mapping between LTM individuals and LTM environments.

2.19.2 Representing the shape of three-dimensional objects

Recall from Section **??** that the map of places in a regular locomotion environment is defined by the system of boundaries: both opposing boundaries and adjacent boundaries.

In this section I will extend this idea to encompass convex environments. The key proposal is that

The idea here is that

My main proposal is that the shape of three-dimensional objects is defined in motor terms.

Here, I want to refer back to the idea that effectors can be locomotors (see Section **??**). The stuff that should go in here is all the stuff about the 'haptic interface revisited' from Part 2.

2.19.3 Representing the configuration of articulated objects

PPs can describe the location of an object, but also its *configuration*. For instance, we can say *John curled into a ball*. I propose that the configuration of an articulated object is represented by the same system that represents the topography of an environment, or at least by an analogous system.

[Then give the model.]

2.19.4 Representing an object's parts or possessions

This is the most obvious one. Here I introduce the idea that parts of an object are **indexed** to that object. So that when it moves, so does the part. This topic connects to the topic of object location memory.

The models of 3D shape and configuration outlined in Sections **??** and **??** are integral to the account of object parts proposed in the current section. The model of 3D shape is important, because it defines a map of places within any arbitrary three-dimensional object. The places within a desk are different from the places within a head: we would not want to use the same map of places in either case. The model of configuration is important in that it provides a way to represent a map of places within an articulated object which is stable over changes in the configuration of this object. For instance, say a flexible belt has a buckle at one end. The buckle is a part of the belt, and must be indexed to a place within it—but this place must be defined within a topography of places that's invariant to changes in the belt's configuration.

2.19.5 Groups as environments

[This shouldn't *just* be about LTM.]

Recall from Section **??** that an LTM individual is an LTM unit that 'points' to a rich property complex. In the discussion in Section **??** I didn't address the issue of how individuals and groups are distinguished in LTM representations. The concept of LTM environments just introduced allows a proposal to be made. What I propose is that a group of objects of a given type is represented in LTM *by an LTM environment* rather than an LTM individual.

The account I propose references the model of visual classification of groups outlined in Section **??**. In this model, when the observer attends to a stimulus comprising a homogeneous group of objects, he can choose to categorise the form *of the group* or the form collectively possessed by the *elements* of the group. In the former case, he is identifying the 'global form' of the stimulus; in the latter, he is identifying its 'local form'. In the model introduced in Section **??**, selecting to classify the local or global form of the stimulus is a matter of selecting the appropriate *classification scale*: if the observer establishes the 'default' classification scale associated with a stimulus of the given size, he will identify its global form, while if he establishes a 'finer-than-default' scale, he will establish its local form. In the model introduced in Section **??**, this perceptual apparatus provides the basis for an account of how the distinction between singular and plural individuals is made. If the default classification scale is selected, any output from the classifier is constrained to represent the type of a single individual, while if a finer-than-default scale is selected, the classifier's output is constrained to represent the type of the individuals in the group. For example, assume the observer is attending to a group of soldiers, standing in a line. If he establishes the default classification scale, the classification system will respond with 'line', or some other type that represents the global form of the group, while it he establishes a finer-than-default scale, the classifier will respond with 'soldier'. My suggestion below will trade on the proposal that by establishing a finer-than-default classification scale, the observer is attending to the group of soldiers *as an environment*. (I didn't refer to it in these terms in Section **??**, because the relevant concept of environment hadn't been introduced.)

At issue is how a plural individual like a 'group of soldiers' is represented in LTM. In the acccount given in Section **??**, an LTM individual always represents a singular object: in this case an entity which has the shape of a line. At issue is how a plural individual like a group of soldiers is represented. Given the suggestion in Section 2.19.1 that there's a 1:1 relationship between LTM individuals and LTM environments, I now propose that the plural individual 'soliders' is represented by the LTM environment that's associated 1:1 with an LTM individual. Just as an LTM individual can be linked to a type or property complex, so can an LTM environment: when it is, this property complex represents the type of the individuals that are in this environment.

Of course, we also need to represent the *number* of individuals in the environment. Or at least, the numerosity of the group of individuals. Here I have another suggestion—namely that the cardinality and/or numerosity of a group of objects is stored in the topological structure of the map of 'places' within the group. Recall from Section **??** that any environment representation is associated with a map of places. For a normal environment like a room, the topology of this map is given by the configuration of boundaries in the environment. A 'group' environment must also be associated with a map of places—but here, I propose that the topology of this map is largely defined by the size of the group.

[This section still has to be completed.]

2.19.6 Agents as environments

Since agents are physical objects, agents can also be attentionally established as environments. It's well known that there are maps of places defined in relation to an agent's motor effectors: for instance, hand-centred maps, arm-centred maps and so on. The key question is: how are these maps set up?

A couple of points to note here. First, agents are obviously highly articulated objects. So the methods for representing places within an articulated object will be relevant. Second, motor systems are extremely hierarchical. So the concept of sub-environments will be useful too (see Section **??**).

Note there's an interesting duality when it comes to representing parts of agents within the environment/place system. On the one hand, effectors can be represented as agents in their own right, which can navigate within their own external environments (see Section 5.2). On ther other hand, effectors can be represented as places within the agent's hierarchically structured body (see Section 2.19.6).

2.20 Representing local relationships between environments

2.20.1 Neighbouring environments

2.20.2 Nested environments

2.21 Representing stable support and containment relations

2.21.1 The relationship *X is on Y*

[Here is where I talk about the idea of a support surface within the object being coincident with the (local) support surface at its place within the environment.]

2.21.2 The relationship *X is in Y*

2.22 Noticing a new object

[This is all about existential sentences: the processes involved in identifying the spatial location of a salient object identified in a retinal coordinate system.]

Chapter 3

Revisions to the spatial representation chapter, based on our recurrent SOM model

3.1 Introduction of the recurrent SOM model of place representation

3.2 Allocentric representations of the agent's orientation in his local environment

This is where you describe head direction cells, and Martin's model of them.

3.3 Environment representations

This is where you describe the environment units that contribute static biases on the dynamics of the recurrent SOM.

3.4 A visual system for representing the agent's place

This is where you describe Chang-Joo's function that learns to map from retinal inputs to the agent's current location/orientation. (And optionally also the agent's current environment.)

3.5 Goal locations

This is where you describe the parallel system of units that represent goals, and how they work.

Chapter 4

Haptic representations of objects

4.1 A recurrent SOM for learning an allocentric representation of the hand's location on a surface

This is Hayim's system.

4.2 A model of 3D object shapes derived from haptic exploration: first attempt

The model presented in Section 4.1 assumes a single hand surface navigating around a single object surface. The model I'll present in this section generalises over that model: it controls multiple effector surfaces, navigating round multiple surfaces on a 3D object. As an example, I'll take the case where the agent's hand is navigating around a cube.

A more general conception of a 'stable support state' In the original SOM-based navigation model, there's just a single effector, travelling around a single surface of an object. In the more general conception, we allow multiple effector surfaces to be touching the object at the same time. They can be touching the same surface of the object or different surfaces. The key idea of a stable support state is still the same: it's maintained if pressure over all the combined surfaces is maintained (modulo slip sensations). There are two things that interrupt a stable support state. One is a change to the set of effector surfaces being supported, either by addition or subtraction of a surface, or by splitting or joining of two surfaces. The other is a change, for any effector, in the ratio between its linear and angular velocity. (This diagnoses a gradual move onto a new surface. For instance, if my fingertip is moving along a flat plane, and then enters a smooth plane with uniform curvature, the ratio of its linear to angular velocity changes from changes from ∞ to a finite constant K.)

An important event that interrupts a stable contact state is the object being dropped. I think this event functions like the event of an agent bumping into the walls of his environment (in a 2D locomotion environment) or reaching the edge of a surface (in a scenario where a hand is navigating a single surface). The action that caused the drop to happen is not used as a training input for the 'next action prediction' network.[1]

A more general conception of 'LTM surface' In the original model, the dynamics of the recurrent SOM is modulated by a tonically active representation of the shape of the surface currently being navigated, in an area holding 'LTM surfaces'. (The analogue of Marco's 'scenarios'.) Different surfaces are represented by different localist units in this area; we envisaged discrete updates to the active unit when the hand transitions from one surface to an adjacent one. In the more general model, the tonic inputs that modulate activity in the recurrent SOM denote distinct **stable contact modes**. While the currently engaged effector surfaces travel around their respective surfaces maintaining stable contact, a single tonic unit encodes the mode of contact that is established.

A more general conception of 'locomotion movement' In the original model, with the hand moving in a single surface, a locomotion movement was a movement of the hand in its own coordinate system. Now, a locomotion movement is a coordinated movement of one or more effectors. These movements still navigate the effectors around the surfaces of the object, and are still ultimately defined with reference to the directions in which hand surfaces travel, as sensed by things like skin stretch receptors. This guarantees that they represent movement of the hand in relation to the object. However, we can no longer refer to just one coordinate system to describe these movements. There is one coordinate system for each separate effector. For example, say the thumb and fingers are touching adjacent surfaces of the cube. One possible locomotion movement is a combined movement of the thumb (on its surface) and the fingers (on their surface). To define this movement, we have to make reference to the local coordinate systems of the two effectors: for instance, 'movement of the thumb in the direction thumb-right, and movement of the fingers in the direction fingers-forward'. If a locomotion movement involves more than one hand surface, it can often alter the position of one hand surface *in relation to the other(s)*. If it does so, it involves a change in hand shape.

My suggestion is that we pass a larger vector of inputs into the recurrent SOM. For each effector surface on the hand, we can pass its local movement on its surface, in its own effector-centred coordinate system. (Those effectors that aren't in contact can pass a vector of zeros.) In addition, we can pass an input representing the current hand shape—that is, something that represents the angles between all effector surfaces. This could be expressed in one of two ways. Firstly, it could be be represented as a static state, derived from proprioception. Seconly, it could be represented cumulatively, by passing efferent copies of the motor commands sent to the movable joints in the hand. (This second way is analogous to the way Martin's recurrent SOM for whole-body navigation is updated.) I think I prefer the first way, because I can't see a sensible way of initialising the

[1] Of course it's not quite the same as a 2D locomotion environment, since the agent's current motor state is not maintained: he has to pick the object up again, and start from scratch.

The case where the agent has two hands in contact with the object's surfaces raises some particular considerations. Here, a locomotion movement can involve simultaneous movements of both hands, in ways that maintain stable support states for effectors in each hand. A movement of this kind can alter the position of one hand surface *in relation to the other*. If it does so, it involves a change in the agent's 'pose' (at least, the relative position of his two arms, and possibly torso too). This is another way of thinking about these movements. The agent simply has to move the surface of one hand in a particular direction *in relation to a surface on the other hand*. If you can represent the location/orientation of a surface in one hand in relation to that of a surface in the other hand, you can represent these directions. (Basically, if the agent has learned, for every location/orientation of one hand, what the motor state is that will bring the other hand into stable contact with this hand, that's half the battle.)

A more general conception of 'allocentric location' In the original SOM-based navigation model, activity patterns in the SOM denote the position of the hand, in a coordinate system centred on the particular surface being navigated. In the more general conception, activity patterns denote *combinations* of positions of *multiple* effector surfaces on the surfaces of a 3D object. I'll call each pattern of activity in the SOM an **object place set**, or just a **place set**. For a given object, and a given stable contact mode, the different states the SOM can get into represent the different places on the object's surfaces that the effectors can establish *in this mode*. Thinking in a localist way, each SOM unit can potentially represent the 3D equivalent of 'a place on a surface'—that is, a *set* of places on an object, each of which is occupied by one of the agent's effectors. That is, each SOM unit represents an object place set.

An object place set represents a set of effector positions. Importantly, these positions are to be thought of as positions *on an object*: they represent the positions of effectors, but they are effectors that are in contact with surfaces of an object, so they denote positions *on these surfaces*. Equally importantly, these positions are given in a coordinate system *centred on the object*, rather than a coordinate system centred on the agent.

A more general conception of a reconfiguration action A reconfiguration action is an action that transitions the agent into a new stable contact mode. It involves breaking the stable support state of one or more effectors, and/or establishing a new stable support state for one or more effectors (either the same ones or new ones, it doesn't matter).

Summary: a new model of the geometry of a 3D object We can now think of a 3D object representation—an **LTM object**—as a localist unit that activates (i) a particular set of stable contact modes, and (ii) a particular set of reconfigurations, mapping from one mode/place to another mode/place.

4.2.1 Analysis of learning in the model

In this section, I'll discuss how the model just outlined should work. I'll first consider the case of a single stable contact mode, involving two effectors. Then I'll consider the complete case, where the agent has to learn transitions between stable contact modes.

4.2.1.1 Learning in a single stable contact mode

Two effector surfaces: fingers and thumb Consider a case where the agent has his thumb on one surface of the cube, and his fingers on the adjacent surface: i.e. when thumb and fingers are on opposite sides of an 'edge'. In this situation, stable contact can be maintained by several kinds of movement. One is when the angle between thumb and fingers is maintained, and the whole hand is moved parallel to the edge. Another is when the thumb and fingers are moved in relation to each other, within their own planes. These two types of movement can be combined, so that the whole hand travels along the edge, and the thumb and fingers also move in relation to one another.

There are very well-defined constraints in this stable contact mode. The size of the cube limits the distance the hand can travel along its edge before it reaches a corner. The edge of the cube places a limit on how far the fingers and thumb can navigate towards each other. And the size and flexibility of the hand limit how far apart the fingers and thumb can travel. Obviously there are interactions between these parameters: there's a little autonomy in the positions of the fingers and thumb, so the hand can move a certain amount in a plane perpendicular to the edge, as well as in the direction of the edge. But as a whole, the navigation movements that can occur are quite closely constrained. And so are the sequences of navigation movements that can occur.

Because sequences of movements are constrained, there should be regularities in the inputs received by the recurrent SOM. Remember that units in the SOM become associated with frequently occurring movement sequences, that is, frequently occurring trajectories.[2] In this case, if we encode some momentum into the hand's movements, there will be frequent sequences associated with arrivals at the two corners. At each corner, there will have been many iterations of the 'move in the direction of the edge' command. (With opposite directions for the two edges, obviously.) So SOM units will learn to encode each of these two commonly occurring sequences. These SOM units represent places (of the hand's surfaces) on the cube—specifically, on its edge. Having learned these places, other SOM units can come to represent other places on the edge, defined as the states reached by specific numbers of movements from the corners.

The above navigation routine assumes only movements parallel to the edge. But as already mentioned, movements are also possible in the plane perpendicular to the edge, subject to constraints relating to the hand's size. These movements introduce an *independent* component of variation. It would be nice if the SOM could learn it separately. In any case, there are some places where movements perpendicular to the edge result in

[2] I'll assume a discrete model of movements, with one every time step: so I'll talk about discrete sequences of movements rather than continuous trajectories.

particularly clear patterns: these again occur at the corners. Say the agent is at the 'top' corner of the edge (call it $E1$) and his fingers are at the adjoining edge ($E2$): there's a well-defined trajectory of the fingers, where they move along $E2$ edge (with the thumb 'following at a constant separation'), until the thumb hits $E1$, and can't go further. There's now another well-defined movement in the other direction, which is curtailed at the point when the fingers reach $E1$. These movements define two more 'points'. (Each is actually a set of points.) These points can serve to define various other sets of points, in which the fingers and the thumb move in relation to one another: if you hold the fingers' position constant, the thumb can establish a range of other locations. The extremities of the thumb's trajectory are defined by the maximum distance that can be established between thumb and fingers.

In summary: I think that after exploration, using a progressively expanding repertoire of trajectories, a recurrent SOM can learn all the object contact states associated with any given stable contact mode.

Three effector surfaces: adding the other hand If we have a more complex stable contact mode, with three effector surfaces, I can't see how things would be any different. The mode still places constraints on sequences of navigation movements, and some of these sequences are still strongly associated with particular allocentric object contact states. From these well-defined reference states, intermediate states can be learned.

4.2.1.2 Learning reconfiguration actions

A 3D analogue of environment exploration mode and environment transition mode In Section **??** I introduced a distinction between 'environment exploration mode' and 'environment transition mode'. In the former mode, interruptions to the stable support state are interpreted as prohibitions on navigation, and define the boundaries of a surface; in the latter mode, they are treated as a cue to find a *new* stable support state. (Maybe the agent is directly *in* a new state, but another altnernative is to make some exploratory movement, to actively *seek* a new support state.) In either case, the establishment of a new support state triggers an update of the current environment.

I assume an analogue of these two modes in the 3D case. To avoid overloading the word 'mode' I'll now call them 'regimes': there is an **exploration regime** and **transition regime**. In the 3D case, the exploration regime is for exploring the current 'stable contact mode'. For any given stable contact mode, the exploration regime is maintained until the agent becomes familiar with the current stable contact mode: that is, until the SOM has learned the regularities within this mode (which is analagous to learning representations of places on the object, or more properly, place sets).

Early stages in learning reconfiguration actions The representation of a 3D object comprises many stable contact modes, and also a set of reconfiguration actions. Each reconfiguration action brings about a transition from one stable contact mode to another.

More specifically, it transitions from a place set PS_1 in stable contact mode SCM_1 to a place set PS_2 in stable contact mode SCM_2.

In the first instance, when the agent makes a reconfiguration action and ends up in a new stable contact mode, all he knows is that he's not in the mode he was just in. He knows this because his predictions about what he can do stop being accurate: so he'll activate a new stable contact mode unit (as in Marco's algorithm), and begin to learn about the place sets contact states this new mode.

Mature representations of reconfiguration actions After two stable contact modes SCM_1 and SCM_2 have been well learned, the agent can learn a new way of representing transitions between them. This hinges on a way of *recognising* the new mode, and the new place-set *within* this mode, when a reconfiguration action is performed.

4.2.2 Spatial relationships between objects

One of the things the agent should be able to do with a learned representation of the geometry of a 3D object is to use this geometry to represent particular locations on an object. For instance, if the agent has learned the geometry of a table, and that of a cup, he should be able to specify that the cup is 'on' the table. (Or perhaps 'standing on' or 'lying on' the table.) This requires a number of things—in particular, the identification of a *single* place on a *single* surface of a 3D object (e.g. a single place on the flat surface (the 'top') of the table). I'll first talk about that.

Identifying single places Does the representation learned in the more general SOM-based scheme for learning 3D object geometry allow the representation of a specific place on an object? Yes, I think so. For one thing, the case where a *single* effector has stable contact with the object (on a single surface of the object, obviously) is a special case of the general scheme. So we can use these cases to identify specific locations. However, there may also be spatial relationships that require the more complex representations learned by the new scheme. I'll consider a couple of these.

Firstly, it's important to be able to represent areas of different *sizes* on the tabletop. A cup occupies a region on the tabletop that's roughly the size of a hand. But larger objects occupy areas larger than a hand. If the agent can put two hands on the tabletop, then larger larger areas can be represented.[3]

Secondly, representations of spatial relations between objects might sometimes require references to multiple contacts. For instance when an object leans against a wall, it makes contact with the wall, but also the adjoining floor.[4]

[3] Actually, that still won't be enough in the general case. Perhaps a place will have to be represented by a particular *trajectory* within a stable contact state, rather than by a static set of effector positions. In which case a single effector would be sufficient, but multiple effectors might still be more economical.

[4] Actually, it's not mainly prepositions that require reference to multiple contacts: it's verbs, like *lean*. (*Against* can be flat up against too.) Other verbs that require reference to opposition spaces include *pinch*, *grip*, *hold*, *grasp*...

I think in the end the main point is that some stable contact modes relate just to one surface, being explored with one effector—and in these modes, single places on specific surfaces of an object *can* be represented.

'Sub-environments' within a 3D object In Section **??** I introduced the idea of 'sub-environments'. A sub-environment is a part of an environment—more than a place, but less than the whole environment. Examples of sub-environments are the 'top' of a table, or its 'edge', or its 'side', or its corner. These concepts generalise over particular 3D shapes, identifying *types* of spatial structure that are found in many 3D shapes.

What sorts of structure are there in the generalised SOM-based model that might correspond to sub-environments? One possibility is that they correspond to *specific stable contact modes* (or possibly sets of such modes). In the SOM-based model, these are localist units, so it's quite conceivable for them to be associated with words. This idea requires that there's a single representation of an 'edge' contact mode—or at least a small number of modes of this kind, that have something in common. So to build a complete 3D object representation, you have to assemble a set of sub-environments together via a particular set of transitions. That sounds vaguely doable.

4.2.3 Transitions between whole objects

Assume that the agent has learned a 3D representation for two objects: a table and a cup. Assume that the agent is navigating on the table: that is, the active object is the table LTM object, which makes available a set of stable contact modes, and defines a particular set of updates between these modes. Now assume the agent bumps into the cup. The representation of this obstacle to navigation is not expressed within the table LTM object. That is, *it is not a known reconfiguration action between the stable contact modes of the table.* This should allow the agent to infer the presence of an object *on* the table.

At this point, an item is logged in the agent's object location memory: *something is on the table* at this location.[5] The agent can then choose to stay with the table, and leave the object on the table unspecified, or to start exploring the object. This latter choice involves a larger transition, from one whole object to another: it involves 'leaving' the table and 'entering' the cup.

If the agent does decide to transition to the new object, all he knows to begin with is its location. When he starts to explore the object, he might *recognise* the stable contact mode he finds himself in. He can then try some reconfiguration actions, to test particular hypotheses about the shape of the new object. Eventually, if he knows the new object, he will be able to activate the relevant LTM object.

[5] This information will also prompt visual attention to this location—but for the time being I'll assume the agent is unsighted.

4.3 Representing the three axes of a 3D shape

Consider a cup on a tabletop: the agent must be able to specify not only its location, but also its orientation. For instance, it can be 'standing', but also 'lying'; it can also be upside down; and if standing, its handle can be pointing in any horizontal direction.

To represent the relative orientation of the cup in relation to the table, I think the agent has to identify three axes for each object. The 'major axis' is often up and down, though it can also be to do with the longest dimension of the object. The next axis is often a 'front-back' axis. For objects with symmetry, some or all of these axes might be ill-defined.

In the current proposed scheme for representing 3D objects, there has to be a way of identifying the major axis. Recall an LTM object is associated with a set of stable contact modes. I propose that the major axis of an object identifies one of these modes in particular—or perhaps a particular set of these modes. (The modes in question are those which constrain exploration along a particular dimension, I guess.) And likewise, the second axis of the object identifies another stable contact mode, or set of modes.

4.4 Representations of scale

I want to say something about the 3D analogue of grid cells here. Two objects of the same shape but different sizes should be represented by the same LTM object: what changes is the 'unit size' for the locomotion 'steps'. If we store for each LTM object what the normal unit size is, then we can record deviations from this normal size; these deviations would provided good denotations for size adjectives like *big* and *small*, which are defined in relation to the normal size of objects of the given type.

If we can independently assign unit sizes for all the dimensions of an LTM object, we can also give sensible definitions of shape adjectives like *thin*, *fat*, *short*, *tall*, *wide* and so on.

The question is: can you apply a particular 'relative scale' for an object's different dimensions in a way that's stable over *all* the object's stable contact modes? That seems tricky.

Chapter 5

A model of motor actions that update spatial representations

[This chapter needs to be rewritten: in the new model, spatial representations are learned within the system that performs locomotion actions. Parts of this chapter will move to Chapter **??**, I expect.]

5.1 Motor systems for interacting with the environment

5.1.1 The orienting system

The **orienting system** is the motor system through which focal attention is directed at a particular place in the observer's current environment. I've already talked about aspects of the orienting system in the preceding sections. I just mention it here for completeness.

5.1.2 Base motor systems

At any given time, an agent is stably supporting himself in his environment. Sometimes this is because the agent is lying down. But the agent can also support himself more dynamically, using motor routines. For instance, when an agent is standing up, he is balancing on two legs, using a motor system including his torso and arms, as well as his legs and feet. If an agent is on all fours, he is balancing using a different motor system, which involves his legs, arms and torso in different ways. I will call these motor systems **base motor systems**, and I will call the base motor system currently in force the **base motor system**. During development, human agents develop a repertoire of base motor systems, including those involved in standing, balancing on hands-and-knees, and sitting.

[Now talk about centre of gravity, and weight-bearing body parts]

5.1.3 Locomotion systems

[1]

5.1.3.1 Egocentric represesentations involved in the control of locomotion actions

While allocentric representations are important in navigation, egocentric representations are also important—often more important. If an agent can see a landmark, then navigating towards it is relatively easy: its horizontal angle in the agent's visual field provides a direct signal about which way the agent should turn in order to navigate towards it. If an agent can see an obstacle, a similar principle specifies how he should turn to avoid it. There appear to be some specialised visuomotor pathways subserving online control of locomotion actions, which are quite distinct from the visual pathways involved in computing environment-centred representations.

Visuomotor routines subserving locomotion actions mainly involve the analysis of visual texture and motion energy at particular points on the retina (see e.g. Gibson, 1950; Perrone, 1992). Like locomotion actions in general, these routines can be defined in relation to the landmark being used to control navigation, and to the trajectory being pursued in relation to this landmark. In this section I will summarise the main routines.

5.1.3.1.1 Navigation in relation to surfaces in the environment As discussed in Section **??**, the agent can navigate using a surface in his current environment as a landmark. The surface in question might be a path through a garden, or a wall in a corridoor; most generally, it can simply be the omnipresent floor or ground. The most obvious trajectories afforded by a surface are parallel to the surface; i.e. trajectories which keep the agent and the surface at a constant distance and relative orientation. For instance, when an agent travels over the floor of a room, he travels parallel to the floor; when he travels down a corridoor, he travels parallel to the floor and to the walls.

There appear to be special visual routines which allow an agent to monitor his progress along trajectories which are parallel to surfaces.[2] These routines exploit the fact that when the agent follows such trajectories, the surface typically projects a relatively *stationary region* onto the retina. For instance, when an agent is travelling down a corridoor, the retinal regions associated with the two walls of the corridoor will each remain relatively unchanging, despite the agent's own motion.

Within the retinal region associated with a surface, there are several cues which the agent can use to control locomotion. One of these is the **texture gradient** in the region. The optic texture projected by a surface becomes finer the further it is from the agent. The gradient of texture fineness through the region can be used to compute the orientation of the surface (in relation to the agent, naturally). A second cue becomes available when

[1] The material below comes from Section 13.11.4 of the 'Part 2' document. So several references to sections are broken.

[2] References needed here.

the agent starts to move. Within the surface there will be a pattern of **optic flow**, which provides information about the direction of the agent's motion in relation to the surface. To take a simple case: if the surface is flat, we can imagine a notional 'horizon line' defining its limit if infinitely extended. If the centre of expansion of the agent's optic flow field is on this horizon line, he is travelling parallel to the surface. If the centre of expansion is below the horizon, he is moving towards it, and if it is above the line, he is moving away from it. Optic flow therefore allows the agent to adjust his movement to pursue a desired trajectory in relation to a nearby flat surface.

5.1.3.1.2 Navigation in relation to object landmarks An agent's locomotion actions can also use landmarks which are objects situated at 'places' within his current environment. Again, there are different trajectories which can be assumed in relation to these landmarks, which can characterised by specialised visuomotor routines.

Assume an agent is navigating *towards* an object landmark. If he is moving exactly towards it, the retinal projection of the object will be at the centre of an expanding optic flow field. If the retinal projection of the object is to the right of the current centre of optic expansion, his current trajectory will leave it to his left, and he must correct his course by turning to the right; if its projection is to the left of the centre of exansion, he must turn to the right. A circuit implementing these visuomotor relationships will steer the agent to the target object. (I think here you can cite work on 'beacon homing' and 'piloting'—see e.g. Whitlock *et al.*, 2008).

If the landmark object is an obstacle, a different set of visuomotor mappings should be used. The appropriate mappings will depend on the trajectory the agent wants to take in relation to the landmark. If the agent wants to navigate *past* a given object, the object's projection should be maintained at a certain distance to the left or right from the centre of optic expansion. (The appropriate retinal distance will be a function of the object's actual distance from the agent, becoming larger as the object approaches.) If the agent wants to navigate *over* the object, its projection should be maintained at a certain distance *below* the focus of expansion, in a similar way; and if the agent wants to navigate *under* the object, its projection should be maintained at a certain distance *above* the focus of expansion.

The case of navigating *around* an object is interestingly different, since it involves describing a curved trajectory. Here the visuomotor routine involves maintaining the object's projection at a particular distance from the focus of expansion (typically to the left or right of it). However, when the object is at a certain distance from the agent, its projection should be maintained at a constant distance from the focus of expansion. This will force the agent into a curved trajectory around the landmark object. Another visuomotor routine for travelling around an object requires the agent to fixate the object, and move in a direction relative to his angle of fixation (see Land and Furneaux, 1997 for evidence that this fixation-based strategy is used by car drivers steering around corners).

Note that obstacles to be avoided can also be defined as objects which are *themselves* moving towards the agent. From the perspective of egocentric visuomotor routines, objects looming towards the agent will generate similar texture flow patterns, and the appropriate motor response can be expressed as a function of these patterns. As noted in the discussion of the reach-to-grasp action, there appear to be specialised visuomotor pathways for the avoidance of objects moving towards the head (see e.g. the discussion of the VIP-F4 pathway in Section **??**). These routines may also be invoked when the agent needs to navigate around a stationary obstacle, at least at close quarters.

5.1.3.1.3 Visual attention to navigation landmarks It is interesting to note that all of the visuomotor routines discussed above associate landmarks with well-defined *retinal regions*. Sometimes these regions are stationary in the visual field, and contain optic flow patterns. At other times the regions are moving along optic flow lines. In all cases, the routines require *reference* to a particular (moving or stationary) retinal region.

The association of navigation landmarks with retinal regions provides the basis for an interface between the attentional routines involved in navigation and those involved in object classification. Recall from Section **??** (and from the model in Section **??**) that in order to classify an object, it must be the most active point in the saliency map. When we describe a locomotion action, we can identify the landmark, which presumably involves the regular process of object classification. It thus appears that when monitoring a navigation action, we can define the most active point in the saliency map as the point currently being used as the reference for our visuomotor navigation control routines. This correspondence is reminiscent of the correspondence between attention-for-motor-control and attention-for-object-classification found in the reach motor pathway, as described in Sections **??** and **??**.

5.1.3.2 Allocentric representations involved in the control of locomotion actions

The notion of a ‘trajectory in relation to a landmark’ seems to be quite well defined at the level of egocentric visuomotor routines, as described in the preceding section. Is it also defined in an allocentric frame of reference? We have already seen evidence that the hippocampus holds environment-centred representations of locations by themselves (Section **??**) and of trajectories by themselves (Section **??**). We have also seen how a trajectory towards a goal location can be defined in terms of reward gradient climbing (Section **??**) and the maintenance of a direction-to-goal representation (Section **??**). But are there ways of representing trajectories *past* or *around* landmarks in an allocentric frame of reference? It certainly seems important for an agent to have allocentric ways of specifying such trajectories. If the agent is generating an allocentric trajectory to a goal location using only a reward gradient and a direction-to-goal representation, the generated trajectory will lead straight to the goal, ignoring any obstacles in the route. It thus seems necessary that the mechanism which generates allocentric trajectories is influenced by the location of obstacles and boundary surfaces in the environment. (Moreover, the agent needs a way of representing the trajectories followed by *other agents*, when he is watching locomotion actions, rather than experiencing them—see Section **??**.) There is recent evidence that the mapping between egocentric and allocentric representations supporting navigation is effected by posterior parietal cortex (see e.g. Whitlock *et al.*, 2008).

One way to encode obstacles and boundaries is to represent them as locations which contribute *negatively* to the reward gradient (see e.g. Burgess and O'Keefe, 1996; Vieville, 2006). Provided certain conditions are met, trajectories generated in an environment containing several obstacles and a single goal will reach the goal. These conditions relate to the amplitude of the 'dip' in the reward gradient generated by each obstacle, and to the configuration of obstacles in relation to the aget and goal locations. While the gradient associated with the goal locations should extend over the whole environment, the trough associated with each obstacle should be relatively local, so that it only influences the agent's navigation when he is close to it. If there are too many obstacles, it may be that there is no trajectory to the target following a monotonically increasing gradient. However, the possibility of sequential search among the available trajectories goes some way to remedying this problem.

Note that the location of obstacles cannot simply be ignored; these locations have to be represented actively, if negatively. This is again reminiscent of the case of reaching; recall from Section **??** that locations associated with 'distractor objects' in a reach task are actively inhibited (see Tipper *et al.*, 1998). But it is also unlikely that all the potential obstacles in an environment are represented equally prominently. Recall from Section 5.1.3.1.3 that visuomotor obstacle-avoidance routines involve the allocation of *attention* (in an egocentric frame of reference) to the object to be avoided. The external object location function (Section **??**) ensures that the location of this object will receive special prominence in the agent's allocentric representation. Finally, note that reward gradients must have particular shapes for particular trajectory types. For instance, if an agent is navigating *around* a landmark object, the gradient must decrease if the agent approaches too close to it, but also if he moves too far away from it.[3]

How can an appropriate gradient pattern be learned for different trajectories defined in relation to landmark objects? One interesting possibility is that the function which generates the gradient is learned during the agent's experience, using the visuomotor routine currently controlling navigation *as a training signal*. Different routines associated with 'past the landmark', 'around the landmark' and so on will generate different trajectories in relation to the attended landmark. These trajectories will be recorded, as the agent moves from one point to another. The agent can perhaps learn a function which deforms the shape of the reward surface around his current location and that of the attended landmark so that a gradient-climbing mechanism recreates the experienced trajectory. One of the inputs to this function would be the trajectory type, defined as the visuomotor routine which creates the trajectory.

Again, note a correspondence with the model of reaching developed in Chapter **??**. Recall from Section **??** that the agent develops methods for representing reach-to-grasp actions as observed trajectories of his own hand onto a target object, because these tra-

[3] This gradient pattern could be generated by superimposing a a positive peak and a higher-frequency negative peak, both centred on the target object; the so-called 'Mexican hat' surface.

jectories correlate with the motor representations which drive his actual movements. Once these perceptual representations of actions have been learned, the agent can recognise similar actions performed by other agents; these learned correspondences thus constitute the foundation for the ‘mirror system’ for action representation. The proposal in the current section is that allocentric trajectory representations are similarly learned as independent perceptual correlates of the sensorimotor routines which control his own actions. Once learned, these representations can then be deployed to represent the locomotion actions of external agents. However, in the current case, they also have a role in planning the agent’s own locomotion actions, in situations where the agent is unsighted.

Before moving on, it is also interesting to consider the visuomotor routine of navigating parallel to a boundary surface in the environment, as described in Section 5.1.3.1.1. Is there a way of characterising a trajectory parallel to a boundary surface in allocentric terms? Note first that a boundary surface is not an object *in* the agent’s environment—rather, it is what defines the spatial structure of this environment. One interesting possibility is that the state in which the agent is following a boundary surface is one in which the function generating activity in the agent’s ‘boundary cells’ can be trained. The model of boundary cells touched on in Section **??** could possibly be expanded in this direction.

5.1.4 An agent as a collection of surfaces

In order to make generalisations over agents and inanimate physical objects, it’s useful to model an agent’s body as a collection of support surfaces. There are two types of surface. The simplest are those on which the agent can rest stably without any active movement. For instance, an agent can lie on his front or on his back without the need for balancing movements. (He can lie on his side too, provided he adopts a suitable static configuration of his arms and legs. If he moves these, he may roll onto his front or back.) I will call these support surfaces **static support surfaces**. Alongside these, there is a set of **dynamic support surfaces** which require activation of balancing motor routines. These are identical to the agent’s set of possible base motor systems discussed in Section 5.1.2: for instance, the standing/walking system and the all-fours/crawling system. These support surfaces are quite abstract: they each involve activation of ongoing motor routines to maintain balance.

5.1.5 Goal places

Goal places are defined in the same coordinate system as current observer place and current subject place.

The location motor controller takes the current subject place and a goal place and generates a locomotion action that moves the observer towards the goal place.

5.1.6 Goal environments

Recall that LTM environments are convergence zones: there's one for every token environment in the world that the observer is familiar with. LTM knowledge of the topological structure of environments is stored as a graph in which the nodes are LTM environments. In a given current LTM environment, there are pointers to all neighbouring environments. I'll assume these are represented in a medium that holds goal environments. There are two related media: one holding candidate goal environments, one holding a selected goal environment (picked as the winning candidate). When you activate a selected goal environment, it will indicate (a) the place in the current environment where this environment can be accessed, and (b) the reconfiguration action that needs to be done when this place is reached.

Assume that the observer knows how to get to each of the environments that's adjacent (or within) his current environment. I envisage a goal environments medium, that indicates for each LTM environment (a proper LTM one!)

5.1.7 Representations of distant environments

Assume a recursively defined search function: when the observer activates an object, it activates the environment it's indexed to, and this in turn activates the environment it's indexed to, and so on. If intermediate environments are stored in a working memory medium, that behaves like a stack, then he can implement a classical AI search of the graph of connected environments until he finds a path to the goal object from his current environment. Since the observer knows how to get from each environment to each of its adjacent environments, as described in Section 5.1.6, the sequence of environments is enough to support progressive navigation to the goal.

Assume that the observer represents the relation between any two distant environments as a sequence of the intermediate environments.

has to learn sequences of intermediate environments that take him from

5.2 Effector-based locomotion systems

It's also useful to model an agent's motor effectors as locomotors in their own right. For instance, an agent's hands can be thought of as locomotors, travelling through their own environments. The environments they move in can be different in a couple of ways. Firstly they can be smaller. (They can often get to places that a whole person can't get to.) Secondly they can be three-dimensional. (Since hands are controlled by arms, they aren't constrained to move along horizontal surfaces.)

[4]

[4] This topic relates to the topic of causative actions (see Section **??**). As will be discusssed in Section **??**, an agent can do actions that are defined by the external episodes which they cause to happen. An agent can observe an episode in which some external object undergoes or performs a locomotion action: for instance, a ball rolling across a table, or a dog jumping into a box. Using the causative action mechanism, the agent

5.2.1 Goal places and goal environments for the hand

5.3 Spatial updates associated with a reach-to-grasp action

5.3.1 Updates that happen during a reach-to-grasp action

The main idea here:

- When I decide to act, I make myself the current subject (within my current locomotion environment).

- If I now decide to locomote, I'd stay as the current subject, and the action would change my location, as already described in Section **??**.

- But if I choose to attend to an external target object instead, I reconfigure my spatial system to represent the location of this object. (i) I set the current environment to be the environment that this object is in (e.g. the tabletop), rather than my own locomotion environment. (ii) I set the current effector place within this environment to be my hand, and (iii) I set the goal effector place to be the place where the cup is. When the current effector location *equals* the goal effector location—or when I'm close enough to know that it will soon do so, on the current trajectory—I establish the cup as a goal *environment*, defined in motor terms as a set of goal motor configurations of the hand (including opposition spaces) as discussed in Section **??**. I select one of these, and my hand achieves it, in a reconfiguration action, and we have a stable grasp state.

- At this point, the stable grasp state needs to trigger a reindexing operation in the environment/place system. The tricky thing is that I now need to think of the cup as the current object, and of its environment ceasing to be the table and starting to become my hand (a part of me), to represent the fact that I now 'possess' it.

- When my current state becomes identical to my goal motor state, as signalled axiomatically by the stable grasp state, my representation of the current locomoting object (i.e. the current effector) needs to change: it needs to stop being my hand, and start being the cup. This somehow has to be implemented in an operation that not only has *motor* consequences, but also results in the cup being indexed to me as a motor environment. Motor environments are special, in that they can move, but in one sense they're still just environments: if you want to find out where the cup is, the answer is now that *I've got it.*

can learn to do a motor action which results in an external object undergoing a particular locomotion action. Crucially, the 'external object' in question can be one of his own motor effectors: thus he can learn to execute a movement (of the torso, shoulder, elbow and wrist) that results in his hand undergoing a certain trajectory, defined within its own local environment. I will discuss this case more in Section **??**.

- Importantly, to properly represent the manipulated object as an extension of my hand/arm motor system, I have to make reference to the position of my hand (as an object) within the manipulated object (as an environment). This is because I need to represent the distance between my hand and the surfaces of the object that make contact with external surfaces (and the relevant angles between the contact surfaces of my hand and those of the manipulated object) and because I need to represent the inertial contributions of the manipulated object (which are different depending on the point at which it's held).

5.3.2 Representation of the stable grasp state

I have a hierarchical representation of my own body parts: within this there's my trunk, and then my upper arm, forearm and hand. I have to literally *add* the cup to the end of this system. How do I do this, and represent the fact that the cup extends my motor system?

Holding the cup changes two main things. Firstly, the whole hand/arm system has a different weight, which means I have to learn a modified set of motor movements for achieving particular movements of the effector. (In fact, the weight issue depends not just on the object, but on its location and orientation in relation to the hand. If you hold a stick by its end, the torque it applies to the hand is much greater than if you hold it by its middle.)

Secondly, the representations of surfaces defined by touch within the motor system in question have to change. Recall from Section **??** that haptic representations of surfaces are defined from first principles from the notions of stable contact and stable support within the body's various motor systems. These definitions have to change when the observer is holding a solid object, because the hand/arm's somatosensory system system registers contacts *made by the held object* against external surfaces as well as direct contacts on its own surfaces: if the held object is solid, tactile sensations are transmitted through it to the hand's surfaces.

The key observation is that the hand can register stable contact between the object it is holding and an external surface. But this is not registered in a first-order sensation of stable contact. (Obviously this sensation is already present when the hand is holding an object, even when it doesn't contact anything else.) It is registered in a *change* to the sensation of stable contact registered in the opposition surfaces on the hand that are contacting the object (e.g. the thumb and fingers). When I lightly press a square block against a vertical wall, I register a new signal within these opposition surfaces. But unlike direct contact, the orientation of the external surface is not given directly by the angle of these hand surfaces: rather it's given by a *new*, *unpredicted* component of force that acts on these hand surfaces. At the same time, there are various navigation movements of the hand/arm that can be done that allow this new component of force to be stably maintained. And there's a similar notion of resistance to changes in orientation. As just mentioned, these directions of constrained movement and rotation do not relate directly to the orientation of the contact surfaces in the hand. The relevant angle is that of the unpredicted component of force within the hand's contact surfaces, rather than the angle of the contact surfaces themselves. And to determine the *location* of the external surface in relation to the hand surface(s), I have to reconstrue my hand as navigating *within the held object*; the location of the external surface is given by the vector from the hand's current location 'on the object' to the end of the object that's in contact with the external surface.[5] (As soon as I stop slipping the object through my hand, I re-establish it as a part of my body, but with a new orientation.) Finally, when it comes to signals that indicate movement over a contact surface, there's no sensation of slip when it's a held object that's doing the moving, but there's a related notion of *vibration* that can be used to diagnose smooth movement of the held object across the surface.[6]

In summary, stable contact of a held object with an external surface is represented as a set of things. Firstly, a constantly acting new component of force on the hand surfaces currently contacting the object. (By 'new', I mean a component that's not explained by the geometry of the object, e.g. by the opposition space within the object that the hand surfaces are contacting.) Secondly, a new definition of the 'support vector' in relation to which stable navigation commands are given. (And this can change as the position and orientation of the object in relation to the hand varies: the object can rotate or translate.) Thirdly, a definition of the hand-movement vector that would bring the hand directly into contact with the indirectly contacted surface, which indicates the distance of the external surface from the hand. Lastly, a new system for movement detection at the contact surface, that uses vibration instead of slip.

5.3.3 Visual routines for monitoring the movements of manipulated objects

When I'm controlling the actions of a held object, I will monitor the object as if it were an external object. For instance, if I'm controlling movements of a held stick against a wall, I won't attend to the point where my hand contacts the stick, but to the point where the stick contacts the wall.

This is where I talk about motor actions defined by the effects they have on external objects, and the causative alternation. But I have to hold off giving a detailed account of this until I've discussed cognitive representations of causality.

[5] If the observer is holding a pole whose end is touching a wall, then the location of the wall is given by the hand-based vector that moves the hand to the end of the pole, which translated into motor actions of the observer's body may well involve locomoting in the direction pointed by the pole. This provides a way of explaining the extended peripersonal receptive fields of cells in parietal cortex when the agent is holding a pole, see e.g. Iriki *et al.*, 1996.

[6] And often acoustic signals that result from this, see e.g. ??. These are often more noticeable for held objects than for hand surfaces, because the former are often more rigid than the latter, and therefore make more noise when they slip across a surface.

5.4 Motor actions that cause spatial changes in a target object

These will be considered in more detail after cognitive representations of causality are considered in Chapter ??. But there are a few things that can be noted right away.

5.4.1 Causing changes in shape

I already considered how to represent changes to the configuration of flexible or articulated objects (see Section ??). When an agent makes contact with an object, he can learn how to make these changes happen. In this case, having selected the target object as an object, he has to re-establish it as an environment in its own right, with spatial structure.

5.4.2 Causing changes in location

In this case, the agent keeps the target object as an object, but represents it within its own environment.

5.5 Old below

5.6 Locomotion actions

In this section, describe the basic idea that choosing to do a locomotion action happens by activating the environment/place system *after having established an agent*. This contrasts with the scenario where the environment/place system is activated when it's referred to the observer's *current* environment, which corresponds to an existential sentence, or (if memory mode has been enabled) in identifying a newly-jumped-to current environment.

5.6.1 Trajectories

5.6.2 Some examples

5.6.2.1 *John walked to the door*

This is an example where the agent moves within his current environment (the room), and ends up being in the same place (within the room) as the door. Exactly what 'the same place' means will be considered in more detail in Section 5.7.1.

The trajectory involved here

5.6.2.2 Example: *John walked across the room*

5.7 Motor actions that transition between environments

[This focusses on reconfiguration actions, and on the idea that an adjacent environment is represented firstly by its stable support affordances—i.e. that these represent 'goal environments'.

5.7.1 Example: *John climbed onto the table*

5.8 Learning the perceptual functions that establish environment-centred representations

This topic is left until last, because it's hard, and also rounds things off very nicely.

5.9 Old below

5.10 Hierarchical relationships between LTM environments

I'm in my chair, but also in my office, in the Owheo building, in Union St, in Dunedin, in NZ, in the world. I like the idea that all of these LTM environments are active at once, in different 'sub-media' within the LTM environment system. If I get 'out of' my chair, I'm left at a place in my office. If I go 'out of' my office, I'm left at a place in the Owheo building. (Note: there's only one way out of my office: and I arrive at a specific place in the 2nd-floor corridor, which is a *neighbouring* place in the Owheo building.) So I'm never in the Owheo building without also being somewhere more specific: a corridor, a flight of stairs, etc. This makes me think about navigation environments more generally.

The Owheo building is made up of a set of **sub-environments**. Each of these has a map of places in it, but importantly the sub-environments are also linked to each other, representing the spatial structure of the Owheo building as a kind of *graph*. The arcs all look like this: if I'm in Corridor C_1 and I'm in place P_1 with orientation O_1, then I enter the adjacent environment Staircase S_1, where I have orientation O_2. Note, in each sub-environment there really are places: these are small enough that I as an agent have freedom to move quite freely in two dimensions (barring obstacles and boundaries).

There are also special places in certain sub-environments (e.g. the downstairs lobby) where I can leave the whole Owheo building. These can be represented with similar arcs. Actually, maybe these places are directly represented as places within the Owheo building. Thus: I come down the main stairs (a sub-environment of the Owheo building) and I arrive in the downstairs lobby (a neighbouring sub-environment of the Owheo building). At a certain place in that environment is the door *of the Owheo building*. If I go to that place, I've left *all* sub-environments of the building, and I'm just at a *place* in the building. At this place, I suggest I activate the environment that the Owheo building is in: namely Union St. Union St has some sub-environments of its own: the Owheo building is one of them; the southern pavement is another. Activating the Union St LTM environment activates a set of map arcs just like activating the Owheo building does: one of these arcs is: if I'm in Owheo building at place P_1 with orientation O_1 and I go forward, then I get onto the South pavement, with orientation O_2. The South pavement of Union St is an environment with places.

I'll distinguish between **topographical** and **topological** information associated with an LTM environment. Topological information takes the form of arc statements: if I'm in sub-environment $SubE_1$ at place/orientation P_1/O_1, and I do action A_1 then I get to sub-environment $SubE_2$ at place/orientation P_2/O_2. Topographical information is expressed in a structured set of points: at each point you can go in any direction (subject to constraints about boundaries and obstacles) and you can define trajectories over these points.[7] An LTM environment can contain both kinds of information. If you're at place/orientation P_1/O_1 there may be directions you can go that just get you to another place/orientation in that same environment—but there can also be a $P_i/O_i/A_i$ that take you into a sub-environment, or into the super-environment, or into a neighbouring environment.

Now I think I'm ready to define how hierarchical relations between LTM environments work.

I assume there can be an **active** LTM environment at several spatial scales simultaneously. (At most one at each scale.) But at any time, only one LTM environment can be **activated**. When an LTM environment is activated, it defines the spatial relationships in the topological/topographical map. (Which I'll call the 'map' for short.)

I'm not sure the above is quite right. An LTM environment is defined as a structure of places, but also a *collection of sub-environments*, which are LTM environments in their own right. If we assume that environments are represented twice, once within the set of candidate LTM environments (where many can be active at once) and once within the set of current LTM environments (where exactly one can be selected), then possibly the relationship between an LTM environment and its sub-environments could be represented by linking a single environment in the *current LTM environment* medium to a *set* of environments in the *candidate LTM environments* medium. So when you establish a new current LTM environment, you also activate an associated set of candidate LTM environments. I also assume a function that says in current environment E_{curr}, you enable a particular bidirectional association between candidate environments and places in the topographical map: so activating a candidate environment will activate a candidate place, and activating (some) places will activate candidate environments.

[7] I'm not sure yet whether the difference between topology and topography is qualitative or quantitative at the level of neural representation.

5.11 Hierarchical relationships between LTM environments: another go

There are two possibilities. One is that there's only one active LTM environment at any time. When I'm in my chair, I'm not in my office: when I get out of my chair, I *reactivate* my office as an environment, and a particular place in this environment. When I'm in my office, I'm not in the Owheo building. (At least not explicitly.) When I leave my office, I *reactivate* the Owheo building as my environment. This doesn't seem right to me. When I'm in a sub-environment in the Owheo building (e.g. a corridor) I can also imagine my position in a topographical map of the whole building. (Taxi drivers have to be good at that, to find an efficient route from A to B.) When I'm in my chair in my office, I can certainly picture myself at a place in my office.

The other possibility is that a notion of spatial scales applies, and that at each spatial scale I can activate (at most) one LTM environment. But still, my *attention* will be focussed at any given time on my environment *at a particular spatial scale*. So, if I shift *in my chair*, during the time that this is planned and experienced, I'll be attending to the chair as my environment. If I notice someone at the door, I'll be attending to my office as an environment.

So: what does it mean to say that my attention is on a given environment? One possibility is that there are several map media—maybe one per scale. At any time, one of these maps is selected, and controls behaviour. (This would be like the competition between a body-centred motor coordinate system for *grab the cup* and an environment-centred coordinate system for *go to the door*. Those are clearly conceived of as different whole maps, which compete against one another in virtue of their summed activity, or something like that.) Another possibility is that there's just a single map medium—say the hippocampus—and the representations in this medium are deictically referred to the active LTM environment, and can change very rapidly.

One way of deciding between these options is to think about the set of active *candidate* places. Everywhere in my office is an active candidate place. But maybe there are active candidate places *outside* my office too. For instance, in the garden outside, which I can see through my window. Say my current LTM environment is my office, but then my attention is drawn to a salient stimulus in the garden, and the garden becomes the new current LTM environment. Was the garden an active candidate LTM environment before my action of attention?

I think it's helpful to bear in mind that spatial representations (those denoted by PPs) are encoded in *LTM*, not working memory. This pushes me towards the view that there's only one map medium, deictically referred to the current LTM environment. But the question then is: how do I shift to other environments? I can (i) go into a sub-environment; (ii) go into the super-environment; (iii) go into an adjacent environment. I think that's it. So: how are these operations implemented? I think that's the big question.

I'll have two cracks at this. One is in Section 5.12; the other is in Section 5.13.

5.12 First go

5.12.1 An environment-transition function

I'm envisaging a fully flat list of LTM environments, and (for the moment) a single 'map' medium holding a structure of places with a parameterisable topography.

I first assume a **environment-structure** function that maps each environment to a structure of places. (This happens through boundary vectors, somehow.)

I now assume an **environment-transition** function that maps an LTM environment E_1 *and* a place in the map P_1 [and a 'transition action', to be explained in a bit] onto a new LTM environment E_2 and new place P_2. The key idea about this is that it's *reversible.* If you're in E_2 at P_2 and you do the inverse of the transition action A (call that A^{-1}), you'll be back at E_1 at P_1.

I want to use this definition of an environment-transition function to do all three sorts of transition: (i) transition between neighbouring environments (sub-environments of some larger environment); (ii) transition from an environment to a sub-environment; and (iii) transition from a sub-environment to its containing environment. The notion of reversible transitions is important in all three cases. (ii) and (iii) are reversals of each other; (i) is reversible in the sense that you can go both ways through a door. Some instances:

5.12.2 Moving between adjacent environments

If I'm in my office (LTM environment E_{office}) at place P_1 next to the (open) door and assume orientation O_1 and do locomotion action A, I get into the corridor (LTM environment $E_{corridor}$) at place P_2 and orientation O_2. Somehow, learning this *also* teaches me how to get from this new state ($E_{corridor}/P_2/O_2$) *back* to the previous state ($E_{office}/P_1/O_1$). Now, how can that be done?

I suggest there are automatic processes in the locomotion system that learn the *inverses* of locomotion actions. Thus, if we walk forwards along some path, we also remember the path, and learn how to walk *back* along this same path. (That could be done by walking backwards, or by turning round and walking forwards; the important thing is the allocentric path, rather than our orientation.) Note that in allocentric terms, the *path* of the action actually doesn't change! It's just the direction the agent travels *along* this path. Maybe this is an important general biological principle: if we go somewhere, it's important to learn how to get back to where we started. In fact, there's very good evidence that rats learn reverse trajectories. When rats pause during a locomotion action, there's evidence that the sequence of place cells activated during locomotion is replayed in reverse, in so-called sharp-wave ripples, at a time-scale commensurate with LTP learning (Foster and Wilson, 2006; Diba and Buzsàki, 2007; Wikenheiser and Redish, 2013). While there are several possible reasons for this, at least one possibility is that the rats are learning the inverse of the locomotion action they just performed. Anyway, I suggest that something similar happens when an agent moves to a new environment.

I'll begin by thinking about how an agent learns the relationship between two environments in a single direction. I'll first consider how this learning that happens in rats. In rats, an agent's orientation relative to the current environment is given by a population of **head direction cells** in a range of (mainly parahippocampal) areas. These cells generate a representation of head direction mostly from visual cues, though path integration is also important (see Yoder *et al.*, 2011 for a summary of evidence). When a rat enters an adjacent environment, place cells in the hippocampus *remap* to encode the boundaries of the new environment (see e.g. Paz *et al.*, 2004); given that head direction cells are computed from perceptual inputs, we might therefore expect that they would also remap. In fact they do when a rat first enters the new environment—but after familiarisation with the transition, they come to maintain their firing when the adjacent environment is entered (see Dudchenko and Zinyuk, 2005). This is even largely true if the two environments are in separate physical rooms, linked by a corridor (see again Yoder *et al*, 2011). Dudchenko and Zinyuk suggest this shows that head-direction cells come to be driven by path integration in preference to perception of the local environment. But another possibility is that the rat is learning to take into account *the angle between the two environments* in its perceptual analysis of the second environment. At the junction between the two environments, it has access to (i) its head direction in the old environment (from path integration) and (ii) its head direction as it would be computed perceptually in the new environment. The difference between these readings can be used to calculate a bias to apply to the perceptually computed head direction reading when entering the new environment, so that it remains stable during the transition to the new environment. (And the inverse bias can be applied when returning in the other direction.)

The idea that large environments are represented by a 'mosaic' of local planar environments with their own reference frames is also advanced by Jeffery *et al.* (2013) in an account of representations of 3D space. They suggest that animals can compute and remember the relation between the local reference frames of adjoining environments; this is as true for 2D environments as for 3D ones.

5.12.3 Going into a nested environment

In one sense, a nested environment is like an adjacent environment. It just happens to be located in the interior of the nesting environment rather than on one of its boundaries.

If I get off the table, I get back into the room. Whereabouts in the room will I end up? If I can see the room, then I can switch between the table and room environments to keep my position updated in both frames of reference. If I can't see the room (say I'm in a tunnel, rather than on a table), I can use dead reckoning to keep my position in the super-environment updated.

5.12.4 Going back into the super-environment

I should be able to re-establish my original place in the super-environment when I re-enter it. If it's shifted, as in the case described above, with a long table or tunnel, I suggest that

I keep it updated by applying dead reckoning periodically in brief switches back to the super-environmnent. So when I come out into the super-environment, I'm at a reasonably accurate location.

An important special case is where I leave the super-environment. If I'm in the downstairs lobby of the Owheo building, and I get to the door, this is the door *of the building*, so I have to leave the sub-environment 'downstairs lobby', and end up not in a neighbouring sub-environment, but simply at a place in the Owheo building (defined topographically). This place is one of the exits, so there's a map in a larger environment describing the environment transition that will happen if I go through this door.

5.13 Second go

5.13.1 Going into a nested environment

So: there's an active LTM environment, defining a map of places. Say it's my office, and the places are the places in my office. Some places *contain objects*: for instance my chair is at one place. It has a particular orientation in relation to the office's coordinate system: it's upright (same as the office), and it's 'pointing' in a particular horizontal direction—which will be some angle in relation to the (fairly arbitrary) 'front' of the office. (Say the 'front' of my office is arbitrarily picked to be the wall adjoining the corridor: that's the wall that faces me as I enter the office. The chair's 'front' will be an angle in relation to this.) The LTM mechanism that indexes the chair to its place in the office will represent the chair's orientation in relation to that of the office. Now: can 'get into' the chair by (i) going to the place it's at, and (ii) orienting myself to the chair's front, and (iii) doing a reconfiguration action (sitting on it). Now I'm *in the chair*. The question is: how has the current LTM environment changed?

When I sit down, my attention is on the chair: specifically, I represent my location in relation to the chair. Is it plausible that possible places I can be at on the chair are represented by *hippocampal place cells*? E.g. one place cell for 'the edge', another for 'the back', one for 'one side'?

Don't forget: there are also representations of location-within-objects elsewhere in the brain, in particular in parietal cortex. The kind of thing that's highlighted in object-centred neglect. If I eat the food on one side of my plate, I'm pretty clearly using a map-based representation of 'locations on the plate', defined relative to the plate.[8] These *parietal* object-centred representations are somewhere between *perceptual* representations and *memory* representations. But I think on the memory side they're *working* memory representations rather than LTM ones. If I was going to encode the location (e.g.) of a computer on a desk in LTM, I'm not sure it'd be in parietal cortex.

Assume for the moment that the map of possible locations in my chair *is* represented

[8] When I say *The carrot was on the plate*, I'm presumably hooking into this representation. So my account of this example has to make reference to an object-centred map of places.

in the hippocampus—and that this map becomes activated when I prepare to sit down,[9] and when I complete the sitting action. I might then want to return 'attentionally' to the room environment. I'll still be in the chair, but I'll want to focus on my position *in the room*, not in the chair. But if need be, my attention can be called back to my position in the chair. The question is: how do I represent my position in these two coordinate systems simultaneously? Or if not simultaneously, then deictically?

Say when I'm standing next to the chair, I create an association between myself (the LTM individual) and my current place in the office (near, or even at, the chair). This represents my location at a particular point in time. Then I get into the chair. I could set things up so that if I activate the LTM environment 'office', I get the place in the office that I'm at, while if I activate the LTM environment 'chair' then I get the place in the chair.

One thing that complicates the above discussion is that the chair is *also* in the office.[10] I can either represent that by requiring the LTM environment 'office' be active *as well as* the LTM environment 'chair', so that the chair is represented as a 'part' of the office. Or I can represent it by having a recurrent operation, that says in order to get to chair, I have to first go to office. I like the second way better, for a few reasons: (i) it allows a wider range of scales of environments (arbitrarily big, really); (ii) it feels more like the context-update operations that happen elsewhere in the SM system.

But if I do it this latter way, then having got into the 'chair' environment, how do I re-activate the 'office' environment? Well, I can think of two ways.

The easiest is for me simply to *attentionally (re-)enter* the office environment. It's still perceptually available, so I can just redeploy my environment-perception modality (through *attention*) so that I get a representation of the boundaries of the environment, and my position within it, just as when I wasn't sitting. This chimes well with my idea that place representations are fully deictic, referred to the currently attended stimulus, which can change from instant to instant. (We know place-cell remapping in the hippocampus is very fast.) Note attentionally re-entering the office environment doesn't put *me* the office—I can't move around in it. If I want to do that, I have to get out of (or get off) the chair. (I have to stand up.)

The other way for me to re-activate the office environment is to *actually* leave the chair environment (rather than just attentionally leave it). I think the way I do this is to do the *inverse* of the reconfiguration action that got me into the chair. It doesn't have to be a literal inverse—getting on and off a table happen through very different actions—but there are learned pairs. In general, the reconfiguration action involves a change of orientation,

[9] Actually when I'm going to sit down, the chair seat is a target, and my *arse* is the effector: in the representation that controls the first part of my reconfiguration action, I'm just reaching with my bum. But the second part is obtaining a stable spatial relationship between me and the chair. (A relationship of support or containment, depending on the kind of chair.) Here I'm not just monitoring my bum: I'm monitoring the stability *of my whole body* within the chair environment. In the way I do when I'm standing. This is a task for a very different neural system, the **postural system** (see e.g. Deliagina *et al.*, 2006).

[10] Just like Union St is *in Dunedin*: it's not anywhere else: it's a sub-environment *of Dunedin*.

5.13.2 Reconfiguration actions

Note: if relationships between coordinate systems for two connected environments are specified in motor terms, then they can perhaps *directly specify a reconfiguration action.* For instance, if I have to step forward to get out of my office and into the corridor, then the reconfiguration action is just 'step forward'. (If there's a boundary that divides my office and the corridor—i.e. the wall that the door is in—then my action can perhaps be defined in the coordinate system of that boundary by itself—and then there's a clear inverse, that works for both environments.)

To get into a chair in my office, I have to work out in advance what the reconfiguration action is. I think I get that by *attentionally* entering the chair first, to find out what transformation is needed. Since this transformation is specified as a motor action, I now know what action I need to carry out.

5.13.2.0.1 Getting onto a table Consider me getting (climbing) onto a table in a room. (i) I navigate *to* the table: i.e. to the place containing the table. I do this using a trajectory computed in my current environment (the room), in which the table is identified by a place: none of its internal spatial structure is visible at this point. (ii) I establish the table as an environment: i.e. as a collection of surfaces. This gives me an affordance-based representation of the table as a set of goal motor states. (iii) I do a movement (a reconfiguration action) and end up re-establishing the table *as the environment I'm in*—or specifically, the environment I'm *on*.

I'm interested in this re-establishment operation, because I think it'll feature in a SM account of PPs (specifically of DP-movement within PPs). The table is represented once as some kind of motor goal, and once as my current location. I need to unpack that idea! It's like the stable grasp state, but rather than the cup being in my hand, *I* am on *the table*. So something has flipped. And it's to do with the difference between a transitive LF structure and an LF structure featuring a trajectory PP.

So: first: there must be some kind of sensation that's the analogue of the tactile sensation of a stable grasp. In the grasp case, the idea is (roughly) that my tactile feeling doesn't change when I move my hand/arm: the target object stays in contact. In the climb-on-table case, I think the constancy, or stability, is assessed in the **balance system**. (This is the system that maintains an agent's centre of mass within its base of support, with minimal postural sway.) I'm going to assume that the balance system delivers a binary signal, 'balanced=true'/'balance=false', and that 'balance=true' has a special role in identifying the concept of *on*. But this is just one case. Say the agent's hand is placed flat on a horizontal surface. There's a similar conception of constancy/stability here: I can't move my hand down or up (without breaking contact); and I can't reorient my hand. And perhaps also: if I relax my arm, my hand registers no slip. I've suggested elsewhere that this type of stability is at the origin of of the relationship of support that external objects have to surfaces. So maybe there's a system that registers this kind of stability, that is also able to issue the signal 'balance=true' to the environment system. Anyway, these are the primary teaching signals.

What I want to think about is how I re-represent the surface of the table when I get onto it and establish a balanced posture. (This could be on hands and knees, or just knees, or on feet, or on one knee, and so on.) One possibility is that I represent the surface of the table as a set of goal motor states, exactly like a transitive object. Then I pick one of these, and do the action that achieves this goal. But this doesn't explain the syntactic difference between an episode in which I grab a cup and an episode in which I get on a table. In the former case, the target object (the cup) ends up indexed to me (I'm the environment, and the target is the object in it). In the latter case, I end up indexed to the target object: it's the environment, and I'm the thing at a place within this environment. I'm sure this relates to the fact that a PP is used to express the climbing/locomotion action (*I climbed onto the table*).

Note that I can't exactly represent the surface of the table that I'm going to get onto as a goal motor state. Because *as a motor state*, the state where I'm standing on the table is identical to the state where I'm standing on the floor. The important thing is that I have to establish a new environment. And since I'm doing an action, I represent it as the environment that *I'm in*, rather than just the new current environment deictically controlling the spatial medium.

Somehow in *I climbed onto the table*, the PP *on(to) the table* represents a transition of environments *for* something, rather than by itself. Whereas in *On the table there was a plate*, the PP *on the table* represents a transition of environments established by me as an observer.

One possibility is that the update operation involves the table environment (and a specific place within it) being established at a point *before the new situation representing the consequent state of the whole episode is encoded in LTM.* So within the WM situation medium—at least, the spatial component of this medium—in the *initial* situation, I'm indexed to the room environment (at a particular place), and in the *new* situation, I'm indexed to the table environment (at a particular place, and with a particular configuration). What would this imply about the environment medium that PP describes? Very simply, that this whole medium represents the location of a *previously selected individual.* The basic idea would be: activate an individual, then activate an environment and (then) a place: that environment/place are understood as being the environment/place *of the previously-selected individual.* In LTM terms, *this* is the object-location-memory-related point. The proposal would be that to author an entry in object location memory, you activate a LTM individual, then you activate a temporal context (and maybe a general spatial context), then you activate a selected LTM environment and place, and *automatically* these things are associated together. That's just how you do it. In experience terms, the deal would be simply that *the currently active* LTM environment (and place) specify where *the currently active* LTM individual is.

Does this selected, active LTM individual have to have any special properties? In *John climbed onto the table*, it's John, obviously; in *John put the cup onto the table*, it's the cup. Why is that? Where is that represented? In the latter case, the cup gets re-represented as a locomotor. Remember there's a choice: are we going to cause the cup to undergo some sort of internal change-of-state (e.g. curling), or are we going to cause it to locomote? If the environment-centred map becomes active, this is because locomotion actions to places in the map are collectively more active (i.e. have higher value) than intransitive actions. That's significant: if the environment-centred map becomes active, then picking a place on the map will result in a (caused) locomotion action. The point is that when I decide to do a causative action on the cup, the active 'place' in my representation of 'the current environment' *stops being 'me', and starts being 'the cup'*. And, for that matter, the active representation of 'the current environment' stops being a representation of *my* environment, and starts being a representation of *the cup's* environment. It could be on a shelf. (In which case I could say *John moved the cup <u>from</u> the shelf <u>to</u> the table.*)

The important thing is that making a change to the current environment *in the middle of a SM routine* has a very different effect from making it at the start of such a routine, or rather at a point *before* it's started. In syntactic terms, the latter cases are signalled by PPs that attach as sentential adverbs; the former cases are signalled by PPs that are part of the subcategorisation frame of verbs.

A key idea in the above analysis is that the process of updating the current *WM situation* interacts in productive ways with the process of updating the current LTM environment. The current WM situation is always updated at the point associated with the 'bottom XP' of the right-branching LF structure of a clause. The original WM situation is associated with the 'top XP' in that LF structure. The bottom XP of one clause is identified with the top XP of the next. These positions are special, in that they are associated with WM situations. Intermediate XPs update various WM structures, but *not the WM situation*: that's only updated once per atomic clause.[11]

Recall from Part 2 Section **??** the idea of the attentional **subject**: who could be the observer, but could also be an external agent. (Or maybe an external object too.) In the account there, the idea was that the observer's spatial medium represented the place of a selected 'subject'—which could either be the observer himself or some other individual. If the observer is selected as subject, this engages a special-purpose circuit that identifies the observer's local environment, and his place within it. If an external individual is selected, *another* special-purpose circuit is activated that identifies the environment of the selected individual (which could be the same as the observer's, but needn't be), and identifies the selected individual's place in this environment. Selection of the subject happens at the very start of a SM routine in the WM episodes system. (When you can either do an action yourself or watch an external agent do an action.) But apparently it can also happen when you activate a causative action. I'll discuss this case specifically below.

5.13.2.0.2 Putting a cup onto a table Say I attend to myself, and then to a cup. By selecting the cup, I've already decided not to do a locomotion action myself. This frees up the environment representation system to represent the environment of the cup. *I assume this happens automatically*, through a biologically hard-wired system. I represent the cup's place within this environment, and then I represent trajectories the cup could

[11] Clauses with nested complement clauses are another matter: in these I think there's an update to the WM situation at the clause boundary signalled by the complementiser *that*.

take. These compete collectively for selection of the environment representation medium. At the same time, a collection of intransitive behaviours compete directly for selection. If the environment medium wins, we know the caused action is going to be some kind of locomotion (i.e. 'go'). The environment representation medium is now *activated.* The next thing that happens is a *trajectory is selected.* Associated with that, a *landmark is selected.* In this case, the landmark is the table. This involves an action of focal attention, which allows the object classification system to generate a representation of the table. Selecting a trajectory means executing it. At the end of the trajectory, the cup is at the same place in the environment as the table is. In practice, I think this means it's above the table, reasonably close to its surface. Now the object undergoes a reconfiguration action. Executing this involves *re-representing the table as an environment.* (And in particular, representing the 'top' of the table as a planar surface.) Once it's represented this way, we can define an action that brings one of the surfaces of the cup into alignment with a surface at a place on the tabletop, and execute this action.

The reason I got started with all this PP stuff was because I wanted to characterise PPs as encoding operations in a system that transitions smoothly from one environment to another one (either a neighbouring one, or a sub-environment: something *local*). I think the above stories show how this happens.

5.14 Learning reconfiguration actions

I got sidetracked talking about what happens when you activate a new environment 'in a motor context', i.e. while in the middle of preparing an action. At this point, I think that the environment representation (and perhaps a selected place *within* this environment too) probably function as a *motor goal*, much like AgrO describes the activation of a motor goal when activated after AgrS (representing an agent). So here's an idea: if the environment representation is updated after an agent has been attended to, the updated representation defines a motor goal. Then an associated movement happens, and if things go well, you achieve your goal state.

How could the relevant action be learned? It could be similar to the cup-grabbing case. You do actions at random; on one occasion you *happen* to change your current environment representation. (Say you happen to *establish a whole new environment.*[12]) This means you can do some learning.

5.14.1 The environment-perception modality again

I assume a *perceptual* modality—nothing to do with memory—that identifies the boundaries of the agent's current environment, and their relationship to one another. My main argument here is as follows. (i) Hippocampal place cells encode a spatial structure defined by the boundaries of the agent's current environment and their relative position. (ii)

[12] A simpler case would be if you establish a new *place* within the *current* environement.

When we're in a new environment, this structure is activated through perceptual information. So (iii) there must be a perceptual modality that identifies the boundaries of the agent's current envionment.

The perceptual modality in question is likely to be implemented, at least in part, in the parahippocampal place area (PPA). This area is known to encode places rather than objects (see Epstein and Kanwisher, 1998 and much subsequent work), and its representations emphasise the spatial structure of boundaries in a perceived place (see e.g. Park *et al.*, 2011).[13] The area encodes both perceptual representations and memory representations of environments (Epstein *et al.*, 2007); for the current discussion, it's the perceptual representations that are most relevant. In sum: when an agent is in an environment, he uses a perceptual modality implemented (in part) in the PPA to generate a representation of this environment as a collection of boundaries with particular spatial configurations; this representation serves to create a spatial map of the environment in the hippocampus proper.

I assume two other perceptual modalities: one that computes the agent's allocentric place in the map of places in the current environment, and one that computes his allocentric orientation in relation to a perceived environment. To the orientation, a constant bias is added, to allow maintenance of a stable representation of orientation across environment boundaries.

5.14.2 What LTM environments are

An LTM environment is a sparse, somewhat localist representation, that represents a particular configuration of boundaries. I assume it literally activates a configuration of boundaries in the PPA. In Damasio's terms, we can think of it as a representation within a convergence zone.

An LTM environment E_1 is also activated by another structure: a combination of (i) *another* active LTM environment E_0, and (ii) a *place* P_0 (within the hippocampus proper). The idea here is: if you're currently in E_0, at P_0, then you are 'at' the place where the environment E_1 is. There are links connecting E_0-plus-P_0 to E_1. But normally these links are gated shut. You have to *enter* E_1—either literally, or attentionally (i.e. in your imagination)—in order to activate E_1 and make it the 'new' current environment. This involves a special *action*: either one of the imagination, or some physical action (a reconfiguration action), that gets you into the environment.

Note that the operation of establishing a new current environment is a recurrent operation. E_0 and E_1 are representations in the same LTM medium, representing environments.

There's one other component to the environment-updating operation, relating to *orientation*. When I'm in E_0 at place P_0 I also have to adopt orientation O_0 in order to get into the new environment. And when I'm in the new environment, I will end up in place P_1, and

[13] In these cases, I'll assume the perceived environment is one you imagine yourself to be 'in'. They're actually presented on a screen, and thus are somewhat too small to be representations of one's actual environment; but there's certainly evidence that the PPA is also activated in immersive virtual reality (see e.g. Bouchard *et al.*, 2009).

I'll have to impose a new orientation bias B_1 on my perceptually computed orientation, so that there's no *change* in my orientation as I transition between environments.

5.14.3 Climbing onto a table again

When I approach a table to climb onto it, I have to establish it as an environment *before I'm in it.* This is like establishing a cup as a goal motor state, as mentioned above. This helps in lots of ways. For one thing, I have to orient myself *in relation to the table* if I want to get onto it. I should orient myself along one of its planar sides: that provides a standard orientation for a stereotypical 'climb onto surface' action.[14] So it's pretty clear that I'm representing the table as an environment *prior* to actually being in it.

How is this done? Do I stop representing myself as being in the room? I doubt it: not yet. (The transition hasn't actually happened yet.) I assume, analogously with the system for executing/learning transitive motor actions, that there are *two* environment representations: (i) the actual environment; and (ii) a selected **goal environment**. (Exactly analogous to the reaching/grasping system, where there's an *actual* motor state and a *goal* motor state.)[15] I suggest that in a certain (perceptual?) situation, the current actual environment is axiomatically copied to the goal environment medium, and a function is learned mapping a visual representation of the environment (as a collection of 3D surfaces) onto this goal environment.

5.14.3.1 Transfer of weight as a perceptual signal for an environment transition

A goal environment is exactly like a real environment in terms of its representation: it's a collection of surfaces. But these aren't surfaces that provide boundaries for navigation: they're surfaces to *reach*, or *get to.* Quite often, the surfaces in the goal environment really are nothing more than reach goals. For instance, you want to get your foot onto the surface of the table. You're 'at' the table, so the table as an object is in your peripersonal space, and its surface can be represented as a goal motor state. (I should do a reach action with my foot so it lands on the tabletop.) When I get my foot there, the table surface can be represented as a reach goal for my *other* foot. When both feet are on it, I can generate a goal motor state in which I'm standing, and balancing, and bring myself into this state (i.e. bring myself upright).

An interesting transition is that in the reconfiguration action which gets me onto the table, my foot starts off just touching/resting on the tabletop; it doesn't *bear my weight.* It's not a *support surface.* When I put one foot onto the table, *I'm* not on the table! (My foot is, but I'm not. I'm still on the floor.) Then I shift my weight onto the table. Here's an idea: *at the point my weight shifts onto the tabletop*—in other words *when I establish a*

[14] If the environment I was entering was a cupboard, there would probably only be one way in: so I'd have to position myself even more precisely in relation to the one way in.

[15] The *goal environment* medium is obviously in the *motor system.*

new support surface in myself—that's when I've established the table environment as my new environment.

When I get onto the table, I think I'm focussing on part of the complete spatial representation. The sides of the table aren't so important; I'm focussing on the horizontal *surface* of the table. I presume what I'm doing is *selecting a support surface* within the goal environment, in the way I might *select* a pair of surfaces (an opposition space) in a cup I'm reaching for. I also have to select a new support surface within *myself*—for instance, one of my feet/legs. Then I have to do two things. First, I have to bring the selected support surface in myself into alignment with the selected surface in the environment. Then I have to *transfer my weight* to the new support surface, so it becomes an *actual* support surface.

What happens here? What's the big deal about transferring my weight? What perceptual modality picks up transfers-of-weight in my body?

5.14.3.2 The base motor system and environment-centred representations

I suggest that at any time, an agent's balance system is *referred* to a particular motor system, called the **base motor system**, defining the agent's current support surface. The postural changes the agent has to make to stay balanced are effected in motor movements that change the position of the agent's centre of mass with relation to the 'base of support' (the body parts that are bearing his weight). For instance, when I'm on my feet, the relevant joints are the ones that move my legs forward, back, and sideways, and that orient my feet forward, back and sideways, and that orient my torso in relation to my legs. (In fact, these movements involve the whole body; but they're normally *overlaid* on top of the body's non-locomotory movements, e.g. reach-to-grasp movements.)

If I'm standing with my feet stationary on the floor, the base motor sytem is the thing that controls my 'swaying'. Note, that even though the motor movements which control swaying are movements of the limbs and torso, their effect is not to move *the limbs*, but rather to move *the agent*. Crucially, they move the agent *in relation to the environment.* It makes sense that when these movements are being done, the agent should *axiomatically* activate a representation of his environment, and a representation of himself as a point *within* this environment, which can move—indeed, which *is* moving.

5.14.3.3 Transfer-of-weight and the body's support surfaces

With the above preliminaries, we can now define a situation in which there's transfer of weight. I assume that at any time there's a function that identifies the agent's base motor system. I'll call this the **base-identification function**. I want to think about the moment when the base motor system identified by this function *changes.* For instance, when the agent drops to his hands and knees, or sits down, or stands up, or lies down. I'll call this a moment of **base motor system change**.

Here's an idea: maybe when we represent our body as a locomotor in the environment, we're representing it as an object with surfaces, *in the same modality as that used to* *represent environments.* Or rather: in a modality which we *learn to map onto* the modality that represents environments.

5.15 Learning reconfiguration actions again

A reconfiguration action can now be defined as an action that changes the currently active base motor system. The rough idea is: you get 'to' an environment (e.g. a table); then you execute a function that represents its surfaces as goal support surfaces; then you select one of these (e.g. the tabletop); then you express this surface as a set of goal motor states (a set of possible locations you can get to) and select one of these; then you select a base motor system as the *upcoming* base motor system (e.g. your hands and feet); then you select an effector in this system (e.g. a leg); then you make the movement that brings the selected effector to the selected goal state, so that it has stability. *Then* you do the crucial environment-transitioning movement: you *put your weight on* the selected effector, and establish the upcoming base motor system as the new *actual* base motor system. At this point you're on the tabletop (on your hands and feet).

I now want to think about how this kind of reconfiguration action (i.e. this kind of environment-transitioning motor operation) is *learned.* I want to think of a mechanism that's somewhat analogous to the mechanism involved in learning a reach-to-grasp action.

In reach-to-grasp, what you learn is a function that maps a visual representation of a target object onto a set of goal arm positions, each associated with a goal wrist/finger position. The learning opportunity is indicated axiomatically by a tactile stimulus.

In transferring your weight to a tabletop, what you learn is (i) a function that maps a visual representation of a support surface onto the goal motor state of an effector that achieves *stable contact* with the surface; and (ii) a **transfer-of-weight** function that *sets up* a new base motor system. This latter function has to *transition* from the *current* base motor system to a *new* base motor system. An action has to be programmed that is—deliberately—*outside* the set of actions supported in the original base motor system, because its role is precisely to *unbalance* the agent. But the action is designed to switch the agent to using a *new* base motor system. So the agent has to learn this function.

In summary, there are two functions to be learned. The first function—the one that maps a visual representation of a collection of surfaces onto a set of candidate goal effector states—is very similar to the function that generates grasp affordances for the hand. Parts of the function are probably identical. Consider an environment which I get into by achieving a stable grasp. For instance a rope, or some rigging, or a horizontal bar in a gym. I reach and grasp this object just as I would reach and grasp a cup. The only difference is that the object is part of an environment: it's immobile, so I can use it to move *myself.* For the moment I'll imagine it's a separate function, computed in a separate bit of circuitry. My main motivation for this is that when I'm looking at a cup as an object to be grasped, I don't evoke the motor state in which my hand is pressed down on the top of the cup, which I would do if I wanted to support myself on it. Generation of this goal motor state is specific to the system that learns environment-transition actions. I'll call the first function the function that generates an attended *surface's* (stable) **support affordances**. As opposed to the function in the reach/grasp system, that generates an attended target *object's* (stable) **grasp affordances**. The second function is the transfer-of-weight function defined above. This function takes a *current* base motor system and a *goal* base motor system, and generates a **reconfiguration action** that *destabilises* the agent in relation to the current base motor system, and establishes balance in the goal base motor system.

5.15.1 Learning the support affordances of a visually perceived surface

The function that learns the support affordances of a visually perceived surface is the function that *defines* surfaces in the haptic system. I've discussed this elsewhere. The axioms for this function are the feeling of support, and the feeling of constraint in hand pose, combined with freedom of hand position. That's what a surface *is*, if defined from first principles in the motor system.

5.15.2 Learning reconfiguration actions

There has to be exploration here, as usual. Say a toddler has balance in base motor system B_1. He will periodically explore a destabilising action A, to see where it takes him. If it establishes a new base motor system B_2, then this motor system is copied to the 'goal base motor system' medium, and a function is learned mapping the original base motor system B_1 and the goal base motor system B_2 onto action A. Simple as that.

5.16 The goal environment revisited

In Section 5.14.3 I introduced the idea that there's a 'goal environment' representation that's distinct from the 'current environment' representation. This goal environment representation is clearly part of the *motor system.* Navigational goals are things that lead to *actions* that update the observer's current environment representation. In this section I want to think a little more about what the goal environment representation is like.

The goal environment representation is often generated from a perceptual representation. But not of the actual environment: of something *within* the environment. The current environment is also represented at the level of perception: e.g. a set of boundaries (that defines the topography of a 2D map of places). The current environment is also active at the level of LTM units: when you recognise what environment you're in, you activate the appropriate LTM environment unit.

Question: is the goal environment representation associated with a unit in the LTM system? I suppose it could be a particular *candidate* LTM environment. Don't forget: in the WM situation medium, a situation is nothing more than a probability distribution over possible WM episodes, and the winning WM episode is *selected*; there's then an update process[16] that creates a new distribution of candidate WM episodes. By analogy, the goal environment representation could be a candidate environment. It would have to be one that *adjoins* the current environment. (So that's the LTM-environment *update* function, analogous to the WM situation update function.[17]) On this analogy, a candidate LTM environment is an environment that's selected *as a goal* by an agent, in the same way that a selected candidate WM episode is one that the agent would like to carry out. I like that idea.

If a candidate LTM environment is selected as a goal, what does that mean? In the case of a selected candidate WM episode, this selection is really just a bias on perceptual and motor processes. There may be things that occur bottom-up that override it. In the case of a goal environment, all other things being equal, maybe the selected LTM environment will bias the observer's attention towards a *place* in his *current* environment. Actually it's just a bias: many places may get some degree of top-down support. This has to be combined with bottom-up information (e.g. about how reachable a place is).

Remember that not all locomotion actions are ones resulting in the trajector getting into a new environment. If I'm in a room and I walk 'across the room', I'm picking and following a trajectory afforded by the room's spatial layout, but I'm *staying in the room.* In this case, there's no LTM environment update at the end of the action. (There's an update of my current *place* in this environment. But that's a more gradual thing anway, I expect: it happens as I'm moving. I guess my place will be updated in the WM *situations* system at the end of the action, but there won't be an update of the current LTM environment.[18]) Actually—see the section on sub-environments below—I think that if you move 'to a corner of' a room, you *can* update the LTM environment. An LTM environment is a collection of boundaries. If you're in a corner, then there are two (adjacent) active boundaries; this says a lot about where you are.

[16] The update process is certainly informed by the current LTM environmnent, naturally! The effects of actions are different in different places; also different actions have different values in different places, so the function that learns the value of an action outcome also certainly takes input from the current LTM environment representation.

[17] Note there's also an *LTM* situation. I haven't been focussing on it, but it's definitely there. So I could equally well talk about LTM situation updates. And now think: an LTM situation is a sparse representation that's associated with a big distribution of possible next situations. It's also associated with one *actual* next situation: that's the big difference. The past can't be rewritten. (But somehow we can talk about counterfactual conditionals.)

[18] Actually it's more subtle than that. If by 'across the room' I mean '*to the other side of* the room', then there probably *will* be a LTM environment update. In this case I end up *at a specified side S of the room.* This is interesting in itself: it's a special case of nested environments. The side of the room is an environment in the room. I guess it's an environment created by a *boundary* of the room. I'll take up this issue in a section below called 'Sub-environments'.

5.16.1 The role of sub-environments in representing *through* and *around*

Even if John walks *through the room*, I think what's happening is that a pair of opposing boundaries are being picked, to structure the direction John is navigating in.

I think if John walks *around the room*, he's structuring the boundaries in the room in a different way, chaining through *adjacent* boundaries, moving parallel to these boundaries, and therefore keeping himself *within* the room. (Moving around a landmark in the middle of a room probably has a similar chaining effect on the set of boundaries that are emphasised.)

5.17 Asides

5.17.1 Issues related to transfer-of-weight

5.17.1.0.1 Transfer of weight in reaching-to-grasp Note the concept of weight transfer is relevant to reaching-to-grasp an object as well. When an agent lifts an object, its weight is transferred to the agent. At this point, the agent must make *postural changes*, to compensate. There's a large literature on this.

5.17.1.0.2 Transfer of weight in walking It's interesting that locomotion *within* an environment *also* requires transfers of weight. I suggest that locomotion happens within a *single* base motor system, using paired effectors within this system. For instance, the legs are paired effectors in the 'balanced standing' base motor system. When you've learned to balance on your feet, you're ready to make some adaptations that support locomotion with the feet.

5.17.1.0.3 Leaning and self-movement Even leaning forwards and backwards, with one's feet planted on the ground, or when sitting, causes motion of the self within the environment. This motion is exploited in cues like motion parallax—and babies do it before they can crawl. (I remember Mia and Helen doing it.)

5.17.1.0.4 Targets of reach-to-grasp actions as environments/surfaces

5.17.2 Aside: *John bent the wire into a circle*

In this case the locomotor object doesn't undergo a trajectory; it undergoes a change in state. The interesting thing is that this can also be described with a PP.

Part 2 already noted that configurations of objects can be represented as environments. If the object ends up in pieces, this makes some sense: a plural group has already been thought of as an environment. What about if the object just ends up in a different shape? (Note: we can talk about the shape *of* the object. Or its size, colour, etc. . .)

Well: we know that an object can be reparsed as an environment. When we do so, we identify its spatial structure. So I don't think this is a very big mystery. The one interesting thing is that (i) 'shapes'—i.e. spatial structures of environments—apparently fall into categories themselves; and (ii) these categories are linked to words (like *circle*, *line*, etc). A circle isn't an object in the world; it's the *shape* of an object. This is interesting from a mathematical point of view, since maths deals with the properties of shapes rather than objects. A rather Kantian conception of maths emerges here, maybe.

I can say that a window is 'round' or 'square'; there I think I'm using the adjective system to describe categories of shape. I can also say a window is 'tall'; I think this is another reference to a category of spatial environment structure, but this one has some reference to the 'normal' spatial structure of windows, so it's a bit more complicated.

(See also *John smashed the cup into pieces* and many others.)

5.17.3 Aside: the re-indexing that happens in a cup-grabbing episode

In a cup-grabbing episode, the cup starts off indexed to the environment, and ends up indexed to me. How does that work? There should be an account analogous to the reindexing account just proposed above.

I can say *John grabbed a cup from the shelf*. Or I can say *John took a cup in his hand*. So it's pretty clear the cup is undergoing a locomotion action, and that this *can* be represented in the environment system. However, the more usual way to think of the cup's location after it's been grabbed is in relation to the concept of *possession*. (Q: Where's the cup? A: John has it.) *Have* is a verb, not a preposition. I already have an idea about *have*: when we say *John has X*, what we're doing is attending to John as an environment, then identifying something indexed to John. In a case like *John has a cup*, we might well add ... *in his hand*; so PPs do get in here too. The point is: the hand environment is already part of John.

I'll leave the rest of this story for some other time.

5.17.4 Aside: object-centred coordinate systems

Recall that the preposition *on* is supposed to denote the identification of a contiguity relationship between a support surface and an object, itself represented as a collection of surfaces. Various *different* surfaces of the object could be contiguous with the support surface.[19] I presume the object has its own coordinate system, so I suggest that the relation denoted by *on* also specifies a change in coordinate system: a relation between the intrinsic coordinate system of the support surface and the coordinate system in which the object (as a collection of surfaces) is defined.

[19] These can be identified in language: I can say that the cup is *standing* or *lying* on the table. Or I can say that the cup is *on its side*, and refer *explicitly* to the surface of the obejct that's contiguous with the support environment.

This makes me think of my ideas about affordance-based representations of boundaries in a room (specifically, walls). There's a spatial relation—an *angle*—between the plane of the floor of my room and the plane of one of the walls in the room. I was thinking of ways of defining this angle, and my thought was that it could be defined in body-centred terms: if my feet are in the plane of the floor, and my position is 'at' the wall, and I'm facing away from the wall, then the plane of the wall is defined by the plane of my back.

In the above case we wouldn't want to say that the wall was 'on' the floor, because both surfaces are part of the room. But consider a large wardrobe that one could conceivably lean against. That can be on the floor. Say I was going to get into the wardrobe. That would have to be done differently depending on whether it was on its feet or on its side. I think when I register the orientation of the wardrobe with respect to the room, I'd want to do it in motor terms. For instance, I could describe the relative angle as the angle that *I* would have to rotate through in order to get onto the floor on my hands and knees (and orient my natural 'up' in the same direction as the wardrobe). This would be a useful way of representing the angle, because having rotated through that angle, I can act on the rotated wardrobe in the way I would normally act on it if it were upright.

I also thought about planes on the hand. If the plane of my fingertips is on the surface of my desk and I move my hand forward to touch a box lying on the desk, the plane of the box is defined by the plane of my first finger-joints.

What is the coordinate system of the box? I'll assume objects for which there's a natural 'up', like a cup. For a box, the natural 'up' orientation is where the opening is facing up. There's also often a natural 'front': for instance for people and animals that are symmetrical. Even for inanimate objects there are often one or more natural 'fronts', from which one can be picked arbitrarily: a 'front' in this case is specifically defined as being a *planar side* (see especially Pereira *et al.*, 2010). So: there's a bottom, and a front—and therefore there's also a *back*, and *sides*. So a box has these things. A cup has a bottom and a top. The top is called the rim, since it has other special characteristics. If it has a handle, it also kind of has a front: if you're left-handed, it's the side with the handle on the left (and the side with the handle on the right if you're right-handed). In sum, objects often have their own intrinsic coordinate systems. At least partial ones.

To represent the orientation of a cup on a table, I might also want to use a motor frame of reference. Say a cup is on its side, with the handle sticking up. We need to define the angle between the 3D coordinate system of the table and that of the cup. I could define this in motor terms as follows. Imagine my hand flat against the plane of the table. Now imagine my hand holding the cup so that the cup's intrinsic 'up' aligns with the table's intrinsic 'up'. The rotation of my hand could define the angle between the table and cup coordinate systems.

I'm not completely convinced of the above.

5.18 Attentionally entering an environment

Consider the following sentence:

(5.1) [On the table][there was a dog].

I suggest that this sentence reports two interlinked processes happening in different media.

The environment medium represents *[on the table]* as a SM routine that transitions from the observer's current environment to the implicit 'subject' that's asserted to be on the table. The process through which this subject is identified semantically is represented in the episode medium, and reported linguistically as *[there was a dog]*. I'll consider these individually first, and then consider their relationship.

5.18.1 The SM routine in the environment medium

The routine that runs in the environment medium begins from two starting points. One is the observer's current environment. (The room he's standing in, I guess.) The other is a newly-attended (i.e. salient) 3D location, at which the presence of a **subject** has been detected. The environment medium *classifies* the identified subject as a structure of 3D surfaces. *But this process is not denoted explicitly by any constituent in the PP.* (At least, not by any constituent at the *top* of the PP.) A constraint operative in the environment medium is that the subject is not properly represented until a stable relationship of support/containment between it and the observer's current environment can be established.[20]

5.18.1.1 How a PP denotes processing in the environment medium

Recall my general SM interpretation of XPs: the maximal projection (XP) denotes the **initial context** in which the SM operation is executed; the head (X) denotes the operation itself (as read from a planning medium); the specifier denotes the reafferent sensory consequence of the operation, and the complement denotes the **new context**. In the case of a transitive clause, there's an interesting situation at the bottom XP: this is the consequent state of the reported motor action, but *also* a state which indexes the transitive object, thereby signalling a piece of cross-modal learning.

It's useful to apply this interpretation to the case of PPs. I've been focussing on operations that take the *current* spatial context and deliver a *new* one. The **spatial context** is a tuple: (i) a current LTM environment; (ii) a current subject location/orientation/configuration. In the case of *[On the table][there was a dog]*, I think that the 3D location denoted by *there* acts as the trigger of the environment update operation. The observer selects a 'candidate subject' (the salient 3D location)—an operation that *activates* the environment medium, and thus is not represented *within* the PP.[21] So the operations denoted by the PP are those through which a new environment is established *to which the candidate subject has a recognisable, stable, spatial relationship.*

[20] Obviously there are exceptions to this, where an environment is established *non-perceptually*. For instance, right now I can talk about what happens *in Tonga*, even though I'm not there. I assume that in this case, *in Tonga* is signalling a transition in the observer to some form of **memory mode**. The collection of candidate LTM environments have become active enough that memory mode is entered, and the LTM environment 'Tonga' is top of the list.

[21] In the same way that the subject agreement projection sits outside the VP, maybe.

The current LTM environment starts off being the one the observer is in (i.e. the room, I guess). But the selected subject is not the observer: it's an as-yet-uncategorised point in external space. Note: it's not just uncategorised *semantically*: that's the case in the DP system, and even the episodes system for an existential sentence like this one. It's also uncategorised *spatially*, and that's the crucial point. There's not much to say about the newly-selected subject at the start of the routine: it doesn't even have a place. (I think it *does* get classified as a structure of surfaces in its own right. This makes good sense in relation to the reach/grasp parietal pathways: the object-centred geometry of the dog is represented in the grasp system, and the location of the dog is represented in the reach system. The key point is, presumably, that there's no linguistic interface to the medium that represents the dog as a collection of surfaces. So notionally, it could be the highest XP in a PP—but it's only present at LF.[22]) It has a place in the DP system, and I guess in the episodes system, because it's defined *retinotopically.* But it's invisible in the PP system—or at least, only identifiable by inference *from the presence of a locative PP.* The PP system reports updates to the current spatial context, and there wouldn't be one (at least perceptually) unless the properties of the subject had changed.

So: what we see reported in the PP is a syntactically permitted sequence of LTM environment transitions. We first represent the place of the newly attended subject *as a place in the observer's current environment*—i.e a place in the room. In the case we're considering, the important thing is that the subject *doesn't have a recognised support/containment relationship with the room at this place.* There's no sub-environment of the room (i.e. no single support surface or pair of opposing or adjoining surfaces) that's coincident with a surface (or sub-environment?) within the subject. So automatically, the observer establishes the object that *does* have a stable support relationship with this place: namely the table. That's identified with a normal DP—*the table*—because it's an object in the room. This operation corresponds to a high agreement projection within the PP denoting 'an action of attention to the table' (as a normal object). This gives an opportunity within the PP to pronounce the DP *the table.* (There will be another opportunity later, I presume.) Then the table has to be established as the new current LTM environment. This is a separate operation. Now we consider whether a stable support relation can be established between the dog, as a collection of surfaces, and a surface/subspace within the newly-established *table* environment. In this case, the answer is 'yes', so we can index the dog LTM individual to the table LTM environment, noting its place and its orientation. Place doesn't show up linguistically, but the nature of the support/containment relation does: that surfaces in the head P *on.*

So, what about the complement of the P *on*? This is supposed to signal the newly-established stable spatial context representation (i.e. a tuple of an LTM environment and a subject at a specified place and orientation, with a specified spatial relationship to the place). It's also supposed to constitute some kind of *learning opportunity.* What might that be? What kind of learning is done in the environment medium? I'll consider this in

[22] Maybe this could be Koopman's (2000) highest XP in a locative PP, which I believe has a phonologically empty head?

Section 5.18.2.

5.18.2 Learning in the environment medium

The learning involved in the above example, which is *static*, doesn't involve learning in the locomotion system at all: the observer isn't actively changing his place within his environment, or moving between environments. I guess the learning relates to a *collection* of functions: (i) the function that identifies the spatial structure of the environment; (ii) the function that identifies the spatial structure of the subject; and (iii) the function that identifies the subject's location and orientation/configuration within the environment. These are learned *together*, I guess; my suggestion is that the situation in which this learning takes place is denoted by the bottom XP position in a PP. I'll take a couple of examples: first one where the observer is the subject; then one where an external individual is the subject.

5.18.2.1 The case of the observer as subject

One example of this collection of functions is the set of perceptual/navigational functions that (i) identify the structure of *boundaries* in the observer's own local environment; (ii) identify the observer's own base navigation system (the relevant 'surface' of the observer that's coincident with a surface or surfaces in the environment); and (iii) identify the observer's allocentric *place* and *orientation* within the environment.

My assumption is that these functions must deliver a representation that's *invariant* in a number of respects. I mean that's *axiomatically* invariant. If the observer changes orientation, the representation of the environment as a collection of surfaces should not change, and neither should the representation of his place in the environment. But the representation of his orientation *should* change. If the observer changes place, i.e. locomotes, the representation of the environment as a collection of surfaces should again not change, and neither should the representation of his orientation, but the representation of his place *should* change. In all cases, the representation of the environment does not change.

Now consider how the observer *computes* the function that delivers a stable representation of his environment. Crucially, this function *takes as input* an environment-centred representation of his current location and orientation. (This is essential for delivering stability, since the observer's perceptual representations change radically as a function of his location and orientation.) So when we say that the observer's representation of the environment should not change, even when he changes orientation and/or location, we are defining an invariance that applies to a structure of *composed functions*. The observer has to learn several functions. One is the function that updates his *place* as a (reafferent?) side-effect of a *locomotion action* (e.g. walking, crawling). Another is the function that updates his *orientation* as a side-effect of an *orienting action* (e.g. an eye movement, a head rotation, a torso rotation, a reorientation with the feet). And finally, the function that takes the current orientation and place and (re-)computes the structure of boundaries in the environment.[23]

There are lots of points to make here.

5.18.2.1.1 This is Wiskott territory. In Wiskott's model of the learning of environment-centred representations (see e.g. Wiskott and Sejnowski, 2002), the key idea is that learning a function delivering a stable environment-centred representation happens by imposing the constraint that orientation and locomotion actions should have a minimal effect on this representation. He also specifies that you can learn the place-change function by holding orientation constant, and you can learn the orientation-change function by holding the place constant.

5.18.2.1.2 The above learning story could explain why the lowest complement in a PP is a DP position. In the PP *in the room* (or equivalently, *in the corner of the room* etc), the DP *the room* appears high, in the Spec of an agreement projection, but also low, at the lowest complement position. I think this nicely reflects the invariance constraint that whatever environment representation is established in the initial context, *that same representation* must also (axiomatically) be established in the *new* context.

5.18.2.1.3 Invariance in the presence of orientation/location changes Note that the kind of invariance that's required can tolerate changes in the observer's orientation or location. The whole point is that these changes *change the inputs from which the stable environment representation is computed.* So by requiring invariance of the perceptually computed environment representation at the end of a orientation or locomotion action, we are indirectly providing constraints on the functions that update allocentric orientation and location.

5.18.2.1.4 Some more steps towards a SM interpretation of PPs A PP is a right-branching chain of XPs. If the 'subject' of the PP is the observer, I assume that one of these denotes an operation potentially updating the observer's orientation, one denotes an operation potentially updating the observer's place, and one denotes an operation potentially updating the observer's base motor system. Each operation involves (i) a motor action (an orienting action, a locomotion action or a balance/reconfiguration action), but also (ii) an associated update to *the relevant aspect of* the current environment representation (i.e. current observer place, current observer orientation, current observer configuration with the environment). Each XP describes an operation—the execution of a function—that *carries out* the relevant update. The updated values are passed as input to the function that updates *the whole environment representation.* This function is

[23] I think the functions are composed in the following order: first the function that computes place, then the function that computes orientation, then the function that (re-)computes the structure of boundaries in the environment from perceptual inputs (and the results of the first two functions). But it could also be that the first two functions are computed simultaneously.

described by the lowest XP in the PP. The new perceptually derived environment representation should (axiomatically) be *the same as* the one generated by the functions. This allows supervised *learning* of all the composed functions, through some process akin to back-propagation, in which an error term associated with the output of the last function (representing how it differs from the required invariant representation) is propagated back to earlier functions.

5.18.2.1.5 The operation of establishing a new LTM environment What's the role of the lowest XP in the PP? This computes 'the whole environment representation'. Why does it do that again? Why isn't this environment representation computed *continuously*? Maybe it is, in fact. But I think there's also a role for an operation at the very end of a PP that carries out a wholesale recomputation of the current environment. That's because *the environment might have changed.*

Let's say the observer steps from his office (through a door) into the corridor. His perceptually-derived representation of his current environment will change radically. In this case, we precisely *don't* want to enforce invariance in the representation being delivered. Instead, we want the constraint to run in the opposite direction: since the change is so sudden, we want to activate a new LTM environment, and compute a new observer location and orientation within this new environment, and learn an LTM transition function.

There's only one invariance that's maintained over the transition, as far as I can tell, and that's in orientation. We maintain that by adding a **bias** (a constant angle) to the orientation computed perceptually.

5.18.2.1.6 Reconsidering cases where there's no environment transition The above considerations lead me to rethink what the circumstances are in which the functions updating location and orientation and the function computing the boundaries of the current environment are learned. I'm wondering whether these are learned *all the time*, and consequently are not reported in language (which focusses on discrete updates). I'm not continuously noticing that I'm in the room, even when I move within it or reorient myself within it.

The unnoticed balance/orientation/locomotion actions are certainly something that Damasio talks about in his model of consciousness; they're important in that model. (E.g. when I apprehend an object, I have to actively saccade to it.) But I'm not sure that the self that becomes conscious here features in any linguistic expressions. (Unless the observer happens to establish himself as an agent and then transition in a nonstandard way to external perception mode, as in *I looked at the cup*, *I saw the cup*.)

Maybe these operations feature in a sentence that reports the observer's location: e.g. *I was in a long corridor*, or perhaps better *I found myself in a long corridor*. In this case, of course, there's no locomotion action, but there's a balance action. (Note I could say *I stood in a long corridor*.) I think

My guess is that *in between* two episodes happening, there's *constant updating* of the observer's balance, orientation and location in his current LTM environment. No pause in this. But that nothing happens at the level of SM routines, because no *changes* are registered. (And that's precisely the point.)

I also suggest that when an episode is experienced that doesn't require a PP to be reported (e.g. the observer grabs a cup, an observed external agent grabs a cup, the cup is perceived to be big. . .), the observer's balance, orientation and location are again being updated in real time. This is an interesting idea, because the *discrete* SM steps taking place in the episode medium proceed in parallel with the *continuous* SM operations in the environment medium, but only the former steps generate linguistic side-effects and discrete updates to LTM.

5.18.2.1.7 The balance/orientation/locomotion system and evolution The idea that the observer's balance, orientation and location are changing constantly chimes well with the fact that these mechanisms are evolutionarily pretty old (at least in mammalian evolution)—for instance, rats have them. So it makes sense that they're mechanisms that other mechanisms are built *on top of.*

5.18.2.1.8 Unconscious movements If I happen to attend to myself as an agent, I can represent orientation actions (e.g. looking at a dog, peering behind an obstacle) or locomotion actions (e.g. walking forward) or reconfiguration actions (e.g. leaning against a windowsill, standing up), and I can report these things linguistically. But the above proposal means that these actions can also be done while some *other* episode is being monitored. For instance, if I observe an episode in my office in which John grabs a cup, and I have to move myself to see what the target of John's action is, then apparently I don't *notice* my locomotion action. It's still represented, just as before. But it doesn't get represented *as a WM episode.* What does that mean? Just that (i) it doesn't get remembered as a unit in episodic LTM; (ii) it can't be described in language.

This may help explain a question that's often raised about the mirror system hypothesis. How can a tennis player run across the court while watching his opponent doing some other action? In response, we might suggest that only the opponent's action is represented as a WM episode; the player's own action may have been *initiated* within a WM episode, but while he's running, his location and orientation are updated at a level below that that's registered by the WM episodes system.

5.18.2.1.9 Locomotion and balance verbs and existential sentences Existential sentences normally use the copula, but can also use various alternatives—for instance *Against the fence [there] leaned a pole*, *Into the room [there] strode a man.* It's significant that these motor verbs denote actions relating to the base motor system. That needs to be fitted into the story somehow.

5.18.2.2 The case of an external individual as subject

Now consider the case where the observer registers there's a dog *on the table.* As discussed above, the 'subject' in this case is an external individual—the dog.

As discussed in Part 2 Section **??**, the observer has to maintain *two* representations of current place, current orientation, current support surface. One is *hardwired to represent himself.* (That's the current *observer* place, orientation, support surface.) The other is able to represent *either* the observer—through some kind of copy operation—*or* an external individual. (That's the current *subject* place, orientation, support surface.)

What the current subject location represents depends on whether the observer establishes action execution mode or external perception mode. In action execution mode, they are simply 'pointers' to the current observer place, orientation, support surface. In external perception mode, these representations are computed by the **external object location function**, which represents the place, orientation and support surface *of the currently attended object*. In this mode, the observer has the ability to establish a new environment *attentionally*, rather than by actual locomotion. The basic idea, as already discussed above, is that the observer makes a salient 3D location the place of the current subject, then classifies the subject as a structure of surfaces, then sees if a stable spatial relation can be established between the current subject and his actual environment at this location. If it can't be, then the observer has to *update the current subject environment.* This involves classifying the object that *is* stably in this place in his current environment (in our example the table) *as a collection of surfaces*,[24] then seeing if the current subject has a stable support relationship with any of these—which it does. At this point, the current *observer* environment is still the room, but the current *subject* environment is the table, the current *subject* place is the place *on the table* where the dog is, the current *subject* orientation is the orientation of the dog *in relation to the table*, and the current *subject* support surface is some surface on the dog. (Since the dog is animate, this will be one of its base motor systems: either 'standing', 'sitting' or 'lying [on front, back, side]'.)

A final point: note that if the observer now re-establishes action execution mode, his current subject spatial context would be instantly reset to the room. And if he had been moving in the room while noticing there being a dog on the table, he will re-establish himself in the right position and orientation in the room.

5.18.2.2.1 The current subject representation as a Damasio-style second-order representation

I already suggested in Part 2 that the current *observer* place does not interface directly with language. It only does so via the intermediary of the current subject location.

The current subject spatial context representation (including current subject environment/place/orientation/support surface) is a very good candidate for what Damasio calls a 'second-order' representation. It can represent the self, by pointing, or *referring*, to a more evolutionarily primitive representation of the self. But it can also represent other things, by invocation of the evolutionarily more recent external object location function.

I like the idea that the cognitive systems that are referred to the current subject represent SM operations more discretely than those directly referred to the observer's body. In fact, I like the idea that *the cognitive systems that are referred to the current subject*

[24] And as a side-effect, as the semantic object 'table' in the DP system

are those that feature in WM episodes, and in PPs, and that are involved in storing representations in episodic long-term memory. As we know, in language it's equally easy for us to refer to ourselves and to other individuals. It makes sense to propose that the level of indirection that allows us to do this also represents SM operations in the brain more *discretely.* On the other hand, I think that the system that's hardwired to represent the observer's location in his current environment operates continuously, in real time, rather than discretely.

5.18.2.2.2 Aside: first-order and second-order representations of feelings I think the above story for allocentric/spatial representations (of places and environments) is echoed by the story for representations of feelings. In the human brain there are 'first-order' representations of emotions, that are hard-coded to represent the observer: these are the ones in the brainstem. (These are analogous to representations of the current observer place.) Then there are 'second-order' representations of emotions, in the anterior insula. These can represent the observer's emotions *or* an external agent's emotions. (These are analogous to representations of the current *subject* place.)

I like the idea that the second-order representations of feelings that Damasio talks about *also* have the property that they can represent equally the feelings of the observer or of some other agent. And also that they are represented more discretely at the 2nd-order level. This is consistent with several studies indicating that the anterior insula is a mirror area, but the posterior insula is not; see e.g. [citation needed]. And also studies indicating that the anterior insula is under-connected in autism; see e.g. Menon and Udddin (2009).

5.18.2.2.3 2nd-order representations and environment updates Here's another idea which chimes with the above. The observer needs to have a system that allows him to establish a *new* environment representation when he walks out of his office (through the door) and gets into the corridor. Even rats have this: when they enter a new environment, their place cells remap, and their head direction gets tweaked. This requirement may be at the origin of the discrete episode-encoding system. Even a rat has to discretely represent the shift from one environment to another. A rat does this by having a representation of 'the current LTM environment', and a network of LTM environments connected by reconfiguration actions.

5.19 Associations between DPs and places in the environment

[This relates back to the original discussion of *[On the table][there was a dog]* in Section **??**.]

Actually, if there was a stable relation between the new subject and the current environment, we wouldn't get an update of the current LTM environment. We'd get an update of the *current subject place.* And this is a representation that is mapped to something in the DP system (I guess a *retinal* location)—so this update shows up in the DP system.

So by itself, this operation is just reported in a DP, whose first XP (referential D) just identifies a new location; the associated *place* in the environment system is not explicit, because language only interfaces very minimally with the system of environment-centred places.

5.20 Summary: the SM operations needed to climb onto a table

This section summarises the earlier sections discussing climbing onto a table. The focus is on the operations that happen in the environment/place representation medium, which is the one that PPs report on.

5.20.0.2.4 Attention-to-self The first thing that happens is John attends to himself. (This is denoted by AgrSP.) This operation activates a bunch of candidate media that compete. In syntactic terms, the winner of the competition will determine the type of XP that attaches to [Comp,AgrS].

5.20.0.2.5 Decision to do a locomotion action The maps that compete are (i) the map representing John's current LTM environment (the room he's in), and (ii) the map of candidate reach targets in John's perispace. In the case we're considering, the LTM environment map wins. There are two consequences. Firstly, John engages his current base motor system (the walking system). 'Engaging' it means 'deciding to use it to do a locomotion action'. There are various actions defined within the walking system: walking, running, skipping, hopping, and so on. In this case it's not specified which specific action is taken.

In syntactic terms, the winning of the environment-centred map corresponds to two things: firstly, adjunction of a VP headed by an intransitive locomotion action (e.g. 'walk'); and secondly, adjunction of a PP as complement of the V head.

Why is the PP attached second? This is a question I'll have to return to. For the meantime, the important thing is that control is transferred to the place/environment medium.

5.20.0.2.6 Selection of a trajectory and landmark I think within the now-active environment representation, John has to select a **landmark** and a **trajectory**. Are these selected separately, or together? I think together. Even when the whole environment is apparently selected as the landmark (as in e.g. *John walked around/through/across the room*), I think the agent is selecting a particular *boundaries* of the room (defining its topography) as landmarks. These aren't always referred to in the DP system—and they're not in the above examples—but they can be, in cases like *John walked into a corner of the room*, or *John walked to the edge of the cliff*.

Alongside trajectories whose landmarks are selected boundaries of the environment, there are trajectories whose landmarks are *objects* in the environment, occupying particular *places*.

In some cases there may be trajectories that compete both as traversal goals and as landmark-related goals; for instance if I go ‘around’ a pillar, it may be because I’m interested in the pillar, or because I’m interested on what’s in the environment on the other side of the pillar. When we report such a trajectory, since there’s a genuine (place-related) landmark, we’ll probably use it.

Are alternative trajectories selected in parallel, or serially? My guess is that it’s like the reach system: you can evaluate several candidate trajectories in parallel, but if you want to take into account semantics, you have to attend to the landmarks serially and store the values of the associated trajectory/landmark pairs offline somewhere where a winner can later be picked. In terms of the attentional process: I think when you attend to a landmark you’ll evoke both spatial and semantic components of it. The spatial components tell you about the landmark as an environment in its own right (e.g. how easy it would be to get under, or on, or behind etc), while the semantic components generate predictions about reward somehow. I’ll talk about reward a bit more now.

Ultimately, we select whole SM routines because they bring about situations associated with reward. I think the WM situation medium is a good candidate for being the medium responsible for the top-down/semantic aspects of selection here, as for reach actions. In the case of a reach action, there’s a SM routine involving several steps, resulting in a particular consequent state: it’s the consequent state that’s associated with reward, so the steps of the routine are learned instrumentally as a unit, rather than individually. The consequent state is represented in the WM situation medium. The WM situation medium also represents the environment the agent is in. I like the idea that different environments compete to be the environment that the agent is in, and that if one of these wins, the agent locomotes to that environment, and there’s a change of (i.e. an update to) the current LTM environment.

Different environments are associated with different amounts of reward. This is very low-level stuff in some ways: it’s how rats navigate. But it’s probably true of humans too. Why would one environment be particularly good? It may be warmer, more comfortable; it may have good associations, because there’s often food there (or warmth/comfort, for that matter). The key idea is that environments have *intrinsic value*. There are also cases where getting into an environment is a small part of a much bigger plan: for instance when I *go to work*, I have to go to the front door, then down the driveway, then down Pacific St, then through the wood, and so on.[25]

I like the idea that within the WM situation medium, nagivation goals compete on an equal footing with goals defined in body-centred coordinates, such as reach goals. In the case I’m thinking about now, let’s say John evokes a strong representation of a navigation goal of getting onto the table. This is so strong that it out-competes all the candidate

[25] So there are *macros* in the navigation system, just as I suggested there are in the WM episodes system. In fact I discussed these macros in some detail in Part 2 Section **??**.

‘transitive’ goals that are being evoked (which are also being represented as ‘candidate situations’, i.e. *achievable* situations). Each candidate situation is somehow represented as the consequent state of the action that brought it about. I quite like the idea that in the WM situation medium—i.e. the top-down part of the selection mechanism—the *location* of the table isn’t specified; all that’s specified is that it’s good to be on it. The bottom-up part of the selection mechanism—i.e. the bit that’s reported in the place/environment system—represents the location, but not the value.[26]

If that’s right, what does a goal place/environment look like in the WM situation? I think it has to be more than an LTM environment by itself. I think the goal needs to specify a *configuration* to adopt—for instance, John doesn’t want to go *under* the table, he wants to get *on* it. It’s also important that the goal is referred to an agent, or at least an individual: for instance, it’s good for the *coffee pot* to go on the stove, but not good for *me*. (And in a competition, it may be good for *me* to get to the goal location, but bad for my competitor to get to that same location.)

An episode as represented in the candidate WM episodes buffer is a stored sequence of SM operations: establishment of an agent, establishment of a target, motor-action. In the case of a locomotion action, we might have: establish-agent, establish-new-environment-as-place, establish-configuration-with-new-environment. For instance, ‘John, table, on’.[27]

How would this work with an action like putting a cup on a table? Here, the situation with value is the one in which *the cup* is in a particular configuration with the environment. The causative action mechanism should be enough to do the rest.

To me, environments/places feature in the WM situations medium in the context of states like ‘Self [is] on table’, ‘cup [is] on table’. These states are basically PPs ‘plus a subject’.[28]

[26] Or rather, it represents something like a ‘path cost’: how easy it is to get to the place.

[27] Note: the table could be a particular *token* table, if it had particular value, but it could perhaps also be the *type* table. That might be enough to place top-down value on a *particular* table in the agent’s current environment, even if the agent had never been on it.

[28] The subject as usual feels like an ‘external subject’, that’s added through a different kind of operation.

Chapter 6

The syntax of PPs

This is still just a sketch...

There's a large literature on the internal structure of PPs; for a recent review see Gehrke (2008). At a first approximation, there's a pretty good consensus that a PP contains the following projections:

(6.1) $[_{PathP}\ [_{PlaceP}\ [_{DP}]]]$

This is an idea originated by Jackendoff (1973) and adopted and extended by van Riemsdijk (1990), Koopman (2000) and van Riemsdijk and Huygbregts (2001) among many others.[1][2] If this is the case, then climbing 'onto the table' really involves first getting to the table, and then configuring onesself so one is on it. *Onto the table* would then look like this:

(6.2) $[_{PathP}$ to $[_{PlaceP}$ on $[_{DP}$ the table$]]]$

And *onto* would be formed by the head *on* raising to adjoin to the head *to*. And, presumably, there would be raising of the DP *the table* to the Spec of PlaceP and further to the Spec of PathP. I assume [Spec,PathP] denotes a *place* within the newly-established locomotion environment. A nice idea: the operation of 'associating' this place with the LTM individual denoted by *the table* is in fact the operation which creates a record in **object location memory** of the location of the table (in the current situation). (Technically it's just the creation of a 'cross-modal association', as for other cases of DP-movement, but since we're representing object position, which can change over time, we're making associations between an object, that endures in time, and a place P *at a time* T.) I assume [Spec,PlaceP] denotes something like a 'goal place' (i.e. a goal motor state, represented in motor coordinates). The operation of 'associating' this goal motor state with the LTM

[1] I should include at least Jackendoff (1983), van Riemsdijk (2007), Den Dikken (2010), Svenonius (2007), and the papers in Cinque and Rizzi (2010).

[2] Modern analyses of PP-internal structure tend to posit many more functional projections in the structure. All this suggests to me that the SM medium denoted by PPs is one in which *bona fide* sequences of operations can be executed, and stored, and replayed. I won't go into the details here; I'll just stick with the simple structure in 6.2.

individual denoted by *the table* may also have some interesting SM interpretation—but as yet I can't think what.

Semantically, analyses of PPs posit many of the same primitives that are yielded within the SM system. For instance, Wunderlich (1991) defines an 'eigenplace function', which for each individual identifies 'the place that it occupies', and an 'int function', which for each individual identifies 'the set of places inside it'. I'm assuming exactly these kinds of relationship linking the episode and individual SM systems.

6.1 Koopman's analysis of Dutch PPs

6.1.1 An account of the distribution of R-pronouns

Koopman's (2000) analysis focusses on an interesting piece of data from Dutch. The word translating the pronominal PP *there* in Dutch is the inanimate pronoun *er*:

(6.3) Hij heeft er gewoond.
He has there lived.
'He has lived there.'

But *er* can also be used in conjunction with an explicit P, in which case it has more in common with a DP pronoun:

(6.4) er op.
it on.
'on it / on there'

In this context, *er* is called an **R-pronoun**.[3]

Within PPs, *er* has a different distribution from full DPs. It has to appear at a position *above* P: it's ill-formed if it stays in the P complement position.

(6.9) *op er
on it

Full DPs, on the other hand, have to appear in the lower position. [4]

[3] In English we can also say *on there* as well as *on it*—but only when *there* is a deictic, and is stressed. As a pronoun, *there* by itself, or *on it* are preferred.

(6.5) Put the cups on <u>there</u>!

(6.6) John cleared the table and put the cups there.

(6.7) John cleared the table and put the cups on it.

(6.8) ? John cleared the table and put the cups on there.

[4] In English there's something somewhat analogous: in the word *thereon*, the genuinely pronominal *there* appears together with the P *on*, and expresses the same meaning as *on it*—but the pronoun has to move out of its regular complement position.

(6.10) op deze tafel
on this table

(6.11) *deze tafel op
this table on

In the *er op* construction, there's good evidence the R-pronoun raises to a position that's stilll *within the PP*. This is nicely shown in an example adapted from den Dikken (2010), and one from Koopman:

(6.12) [Er op] heb ik dat boek gelegd.
There/it on did I that book put
'I put it <u>on there</u>.'

(6.13) de tafel, [waarop] ik dat boek heb gelegd
the table, whereon I that book did put
the table, on which I put the book

(Note that *waar* is also considered an R-pronoun: in Example 6.13 the fronted R-pronoun is joined morphologically to the preposition.)

However, R-pronouns can also raise *out* of PP. They can land at a position local to the verb (see Example 6.14), or higher, to a position dominating the whole clause (see Example 6.15):

(6.14) Ik heb *er* dat boek *op* gelegd.
I have there/it that book on put

(6.15) *Waar* heb jij dat boek *op* gelegd?
Where have you that book on put
'What did you put that book on?'

(In each case, the preposition is 'stranded'.) Non-R DPs can't raise like this:

(6.16) *Ik heb *de tafel* dat boek op gelegd.

(6.17) **Welke tafel* heb je dat boek op gelegd?
Which table have you that book on put
'With table did you put the book on?'

Koopman argues that *er* uniformly raises to *specifier* positions. In Example 6.12, it raises to the specifier of a functional projection that introduces the PP. Koopman calls the functional projection **PlaceP**: it's a projection that introduces a 'locative head'. (Koopman argues that PPs have a 'left periphery' of functional projections, just like clauses and DPs do: I'll motivate this idea below, in Section 6.1.2, and say more about PlaceP.) In Example 6.14 *er* raises *beyond* [Spec,PlaceP], to the specifier of the VP that takes the PP as an argument. And in Example 6.15 it raises still higher, to [Spec,CP]. The highest specifier position in the PP, in this case [Spec,PlaceP], is seen as an 'escape hatch' for material within PP: if a constituent can reach this high position, it can move out of PP to higher Spec positions in the clause. *Er* can get to [Spec,PlaceP], and can therefore raise further. Even a PP featuring a full DP can also raise to [Spec,PlaceP], and then move *as a whole* to a higher position, either local to the verb (see Example 6.18) or dominating the clause (see Example 6.19):

(6.18) Ik heb *op de tafel* dat boek gelegd.
I have on the table that book put

(6.19) *Op welke tafel* heb jij dat boek gelegd?
On which table have you that book put

This notion of [Spec,PlaceP] as an intermediate landing site links to accounts of successive-cyclic movement of heads: it originates with van Riemsdijk's (1978) analogy with V-inversion. ('English Vs can't raise to the head of CP because they can't reach the head of IP.') In fact, *er* is often analysed as a clitic. (And this analysis looks reasonable for Example 6.14, where *er* is local to the main verb, and Example 6.13, where the R-pronoun *waar* is morphologically part of the pronoun.) But Koopman argues the movement of *er* is XP raising to Spec positions, not head movement.[5]

Koopman's analysis of PPs addresses two key questions. Firstly, what is the relation between *er* as a locative pronoun and *er* as an R-pronoun? They certainly have different meanings—but why are they *homophonous*? Secondly, why can *er* as an R-pronoun move outside the PP, stranding its preposition (see Examples 6.14 and 6.15), while a full DP cannot move like this (see Examples 6.16 and 6.17)?

Koopman's answer to the first question is that the locative pronoun *er* (='there') and the R-pronoun *er* (='it') *both* move to the specifier of PlaceP. Koopman proposes a scheme whereby this specifier position has to have *something* in it: this something can be a PP, or an R-pronoun complement of a PP, but it can't be a full DP complement of a PP. In this scheme, the Spec position to which a pronominal PP raises (see Figure 6.1a) is the same as that to which an R-pronoun raises (see Figure 6.1b)—and this 'structural' commonality explains why they have the same pronunciation. The idea that PPs are required to raise to [Spec,PlaceP] also provides an answer to the second question: why R-pronouns like *er* can strand the P that introduces them. Koopman assumes that a PP containing a *full* DP must also raise wholesale to [Spec,PlaceP] (see Figure 6.1c). In this configuration, the DP *within* the PP is blocked from raising by a general prohibition on raising material out of a 'left branch' of XPs.

6.1.2 The left periphery of prepositional phrases

Koopman's analysis of PPs is actually more complex than is shown in Figure 6.1: she argues that the projection headed by P (i.e. PP) is dominated by *several* functional pro-

[5] In fact, 'R-raising' as it's called seems to fall between two stools: it has something in common with both head-movement and XP-to-Spec movement. van Riemsdijk (1999) calls it a 'hostless clitic'.

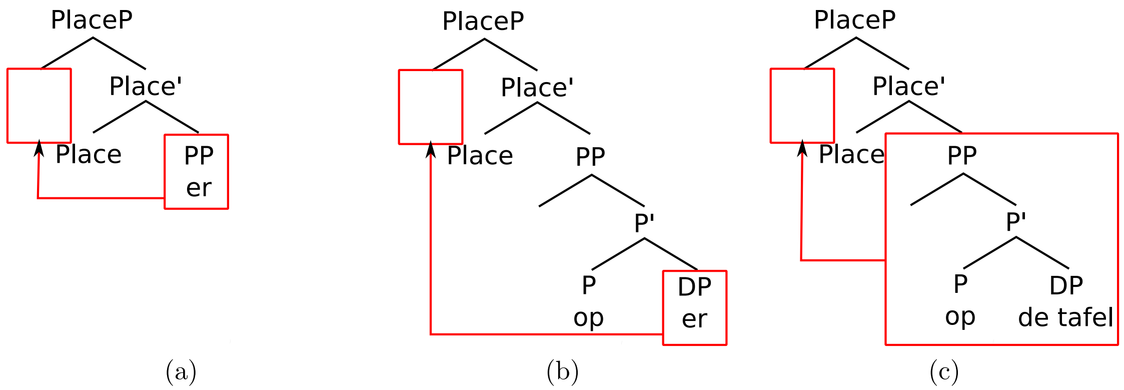


Figure 6.1: (a) A pronominal PP raising to the specifier of the PlaceP introducing PP. (b) An R-pronoun raising to this same position. (c) A PP containing a full DP raising to this same position.

jections, that contribute different kinds of information to the larger prepositional structure, to create the classical right-branching XP structure typical of DPs and clauses. There is a terminological issue here: we want to use the term 'PP' to refer to the XP at the *tail* of the right-branching structure (the one headed by P), but also to the XP at the *head* of the structure (which occupies the position of a PP argument or adjunct in a wider syntactic structure). I'll use the term 'PP' for the 'low' XP headed by P, and the term **prepositional phrase** (or **pp** for short) for the 'high' XP.

Koopman's evidence for the functional projections dominating PP comes from various options for placement of the R-pronoun *er* in relation to modifiers. The basic data relate to degree modifiers, as shown below.[6]

(6.20) [Tien meter *er* naast] heeft Jan gezeten.
Ten meters there next-to has Jan sat

(6.21) [*Er* tien meter naast] heeft Jan gezeten.
There ten meters next-to has Jan sat

In these sentences the prepositional phrase is fronted, so *er* is clearly within this phrase in each case. Assume that degree modifiers appear as specifiers of an XP above PlaceP—Koopman calls this XP **Deg(Place)P**. The *er* that comes after a degree modifier can occupy the specifier of PlaceP—but we have to posit another XP *above* Deg(Place)P, whose specifier can host the er that occurs before a degree modifier. The two positions for *er* are highlighted in Figure 6.2. Koopman calls the XP above Deg(Place)P **C(Place)P**, by analogy with the highest functional projection in a clause (CP). She argues that only a

[6] These examples (and several subsequent ones) are actually from den Dikken's (2010) overview of Koopman's model. den Dikken in fact proposes an extension of Koopman's model, which I will introduce later in the chapter. In fact a lot of den Dikken's analysis was developed earlier (see den Dikken, 2003); I will occasionally refer to that earlier paper, which contains some material not included in the later one.

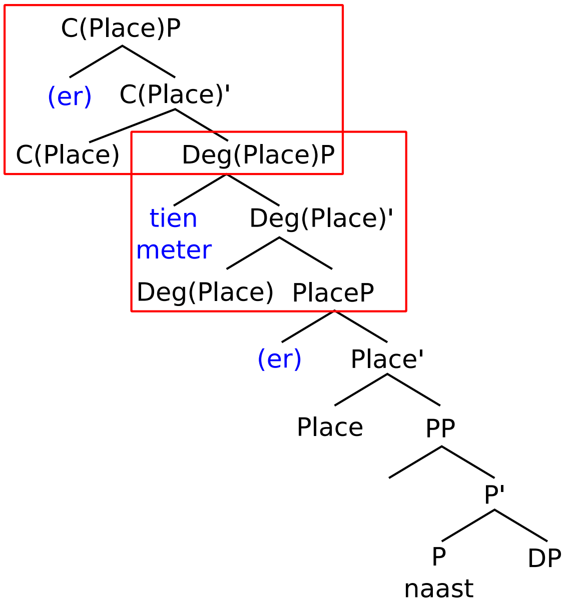


Figure 6.2: The 'left periphery' of a locative PP: C(Place)P and Deg(Place)P

whole C(Place)P can undergo various kinds of movement—for instance, the kind of fronting shown in Example 6.21. It's very clear that smaller XPs within C(Place)P can't front by themselves: if we front PP (with an in-situ full DP, see Example 6.22), or PlaceP (with *er* raised into its specifier, see Example 6.23), the results are 'crashingly bad':

(6.22) *[Naast de deur] heeft Jan tien meter gezeten.
Next-to the door has Jan ten metres sat

(6.23) *[er naast] heeft Jan gezeten.
There next-to has Jan sat ten meters

With this left periphery in place, Koopman must make a further stipulation about head movement within the prepositional phrase. If the P head could raise to the heads of Deg(Place)P and/or C(Place)P, it could appear to the left of *er*, which is never allowed. Koopman stipulates that P can't raise to these higher heads—the highest it can get is to the head of PlaceP. This stipulation does useful work in an account of P incorporation (see Section 6.1.3), and in an account of postposed PPs (see Section 6.1.4.2).

6.1.3 Evidence for Koopman's analysis from P incorporation

Koopman's analysis provides an attractive account of some constraints on a phenomenon called **incorporation**. I'll begin by introducing this phenomenon.

The phenomenon relates to a well-studied structure in Dutch called a **verbal cluster**. In Dutch, as in many Germanic languages, the verb of a finite subordinate clause ('I think that [clause]') or of a nonfinite clausal complement ('I want [clause]') must appear at the end of the clause. Verbal clusters are created when nonfinite clausal complements are nested *within* subordinate clauses. In these circumstances, there are two verbs that must appear 'at the end': the 'outer' verb of the subordinate clause, and the 'inner' verb of the complement clause. If syntax is context-free, we expect that the outer verb will come after the complement clause it introduces—and thus that the inner verb will *precede* the outer one. And indeed, this is the order we find in German. But in Dutch, the inner verb must *follow* the inner verb.

(6.24) Ik geloof dat Jan het boek wil lezen.
I think that Jan the book wants to read
I think that John wants to read the book.

In the resulting structure, notoriously, there are **crossing dependencies**: the outer verb (*wil*) seems to intrude within the structure of the most deeply nested clause ('het boek (. . .) lezen'). If there are nested nonfinite clauses, there can be several intruding verbs.

(6.25) Ik geloof dat Jan Marie het boek wil laten lezen.
I think that Jan Marie the book will let read
'I think that Jan wants to let Marie read the book.'

The sequence of terminal verbs—a finite verb followed by a sequence of nonfinite ones—is called a verbal cluster. The verbal cluster is underlined in Examples 6.24 and 6.25.

Koopman's interest is in a phenomenon called **P incorporation**. This arises if the most deeply nested verb takes a *bare* P complement, to create what in English would be called a 'phrasal verb': for instance *op bellen*, 'to call up' (literally 'up to-call'). In this circumstance, the P can raise to other positions in the verbal cluster: specifically, to positions on the left or the right of the finite verb.

(6.26) dat ik Jan Marie *(op)* heb *(op)* willen laten bellen.
that I Jan Marie (up) have (up) want let call
'that I wanted to let Mary call up John.'

This type of incorporation is not possible if the most deeply nested verb takes a full locative prepositional phrase (e.g. *to sit next-to the door*). To show this, I will first show the structure that *is* permitted, in which the full prepositional phrase raises to the left of the verbal cluster, to the so-called 'central field':

(6.27) dat Jan [$_{Prep_Phrase}$naast de deur] heeft gezeten.
that Jan (next-to) the door has sat
'that John sat next to the door.'

One might think that prior to this raising, the P might be able to incorporate into the verbal cluster by itself, like a bare P. (In this case, the prepositional phrase raised into the central field would be missing a P.) But this is not possible:

(6.28) dat Jan [$_{Prep\_Phrase}$ de deur] *(*naast)* heeft *(*naast)* gezeten.
that Jan the door (*next-to) has (*next-to) sat
'that John sat next to the door.'

However, it is possible for *er* to raise out of such a prepositional phrase, and on to a high position of the sentence, stranding the preposition by itself in the central field:

(6.29) dat zij *er* dit vaasje [$_{Prep\_Phrase}$ op] heeft willen zetten.
that she there this vase up has want put
'that she wanted to put this vase up there.'

Koopman argues that incorporation of a P into the verb cluster is only possible if V *directly* introduces the projection containing P. She suggests that is what happens in bare P complements like *op bellen.* If V introduces a 'full' prepositional phrase, that is, a C(Place)P, as in *[naast de deur] zeten*, the P can't raise to the head that V directly introduces, since it only raises to the head of PlaceP (see the stipulation at the end of Section 6.1.2). However, note that *er* is still able to get to the *specifier* of this C(Place)P, and from there it can raise beyond the prepositional phrase, stranding the preposition, as shown in Example 6.29.

6.1.4 Koopman's account of directional PPs

Directional PPs denote not a static location, but a path, or trajectory. Koopman's basic proposal is that directional PPs introduce a special type of XP called **PathP**, that dominates all of the projections so far introduced for a locative C(Place)P. This idea harks back to Jackendoff's (1983) proposal that directional PPs feature a PathP projection with a PlaceP projection as its complement: but it is expressed in a way that accounts for some of the syntactic intricacies of Dutch PPs. Koopman achieves this with an elegant proposal: that PlaceP can take various different *parts* of a locative C(Place)P. Firstly, it can take a *complete* C(Place)P. Secondly, it can take a *just a PlaceP*, without its 'left periphery'. Finally, it can take *just a PP*. These alternatives respectively account for three distinct types of directional PPs attested in Dutch: **prepositional** directional PPs (see Example 6.30), **postpositional** directional PPs (see Example 6.31) and **circumpositional** directional PPs (see Example 6.32).

(6.30) Hij klimt [in de stoel].
He climbs in(to) the chair
'He climbs into the chair.'

(6.31) Hij klimt [de stoel in].
He climbs the chair in(to)
'He climbs into the chair.'

(6.32) Hij loopt [onder de brug door]. He walks under the bridge through
'He walks under the bridge and through it.'

Koopman's analysis accounts for the internal structure of these alternative types of PP, but also for various constraints on their distribution, and for two key generalisations. Firstly, Dutch *locative* PPs are always *prepositional*. (But prepositional PPs can also be directional—in fact Example 6.30 is a case in point.) Secondly, Dutch *postpositional* PPs are *always* directional.

In the remainder of this section, I will introduce Koopman's analyses of prepositional, postpositional and circumpositional directional PPs.

6.1.4.1 Prepositional directional PPs

In the analysis of C(Place)P introduced above, prepositions always appear before full DPs: this already accounts for the fact that Dutch locative PPs are always prepositional. To account for prepositional *directional* PPs, Koopman suggests that under some circumstances, PathP can take a complete C(Place)P as a complement. Koopman assumes the head of PathP is empty—this means something has to happen to **license** this XP (that is, allow it in the LF structure). One way of licensing PathP would be to move a lower head into it: but we have already stipulated that P can't move into the head of Deg(Place)P or C(Place)P, so it can't raise higher than PlaceP. Another way PathP can be licensed is by being 'incorporated' into the head of the VP that introduces it. A key idea is that individual verbs can specify whether they allow incorporation or not: essentially, this is part of a verb's subcategorisation frame. We can therefore explain fairly idiosyncratic facts, such as the fact that the verb *lopen* (to walk) can't take a directional PP, while the verb *klimmen* (to climb) can take one. Thus, for instance, Example 6.33 doesn't have a directional reading, while Example 6.34 does:

(6.33) Jan liep in de kamer.
Jan walked in the room
'Jan walked (around) inside the room'

(6.34) Jan klom in de boom.
Jan climbed in the tree
'Jan climbed (around) inside the tree' / 'Jan climbed *into* the tree'

We can explain this by stipulating that *klimmen* happens to allow incorporation of an empty PathP, while *lopen* does not.[7] (Both examples have a locative reading, where the PP specifies 'where the action took place', rather than a trajectory. So Example 6.33 has just one reading, while Example 6.34 is ambiguous.) Koopman also stipulates a general subcategorisation constraint, to help account for the distribution of locative PPs: nouns *never* allow incorporation of an empty PathP, which explains why nouns can never take locative prepositional PPs, and thus, for instance, why Example 6.35 only has a locative reading.

[7] It's not clear that this concept of 'incorporation' is quite the same as the one implicated in moving a bare P into a verb cluster (see Section 6.1.3). All that's needed is a device that will license a directional PP complement for some verbs, and not for others.

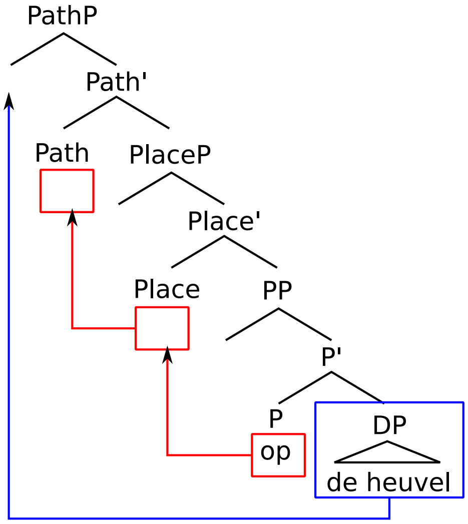


Figure 6.3: Koopman's analysis of a postposed (directional) PP. The P (*op*) raises to PathP to license it; its complement DP (*de heuvel*) raises to [Spec,PathP], so it appears before it.

(6.35) de weg op de heuvel
the road on the hill
'the road [that is] on the hill'

6.1.4.2 Postpositional directional PPs

Dutch nouns can take directional PPs, but they have to be *postpositional*. Example 6.36, for instance, has an unambiguously directional interpretation:

(6.36) de weg de heuvel op
the road the hill on
'the road onto the hill'

There are two things to explain about this example. Firstly, we must explain why the preposition *op* is postposted in this case. Secondly, we must explain how the top-level projection of the directional PP (PathP) can even *appear* in this context. In Section 6.1.4.1 it was stipulated that nouns can never incorporate an empty PathP head—so there must be some other way of licensing PathP in this grammatical context.

Koopman makes a proposal that addresses both of these questions. She suggests that in a postposed PP, PathP takes PlaceP *as its direct complement*, as shown in Figure 6.3. In this case, Koopman argues P can raise *beyond* the head of PlaceP, to the head of PathP. This licenses PathP 'from below', which explains how it can appear in a position where incorporation of the PathP head is systematically disallowed. The presence of PathP also makes available a high Spec position, that the complement DP of P can raise into. Recall from Section 6.1.1 that DP can't raise to [Spec,PlaceP]: only PPs have the 'Place' feature that allows them to appear at this position. But there is no such restriction for

[Spec,PathP], which allows the DP to raise to a position in front of the raised P, giving a postposed P. (Why the DP *has* to raise to [Spec,PathP] is another question: I presume Koopman would say DP 'has a Path feature', or something like that—which is basically just stipulating that it *does* raise.)

Note that postpositional directional PPs can also occur as arguments of verbs—that is, in positions where PathP *can* incorporate. Example 6.31 (repeated below as Example 6.37) is a case in point:

(6.37) Hij klimt [de stoel in].
He climbs the chair in(to)
'He climbs into the chair.'

Koopman accounts for these by saying that a 'filled' PathP head can incorporate to an immediately dominating V just as well as an unfilled one. Note that in structures like this one, the PP can only be interpreted as directional. This fits well with the model, since the postposed P structure requires the presence of the PathP constituent.

6.1.4.3 Circumpositional directional PPs

A circumpositional PP has a two P-like elements, one preceding the DP, the other following it. The example given above, Example 6.32, is repeated below:

(6.38) Hij loopt [onder de brug door]. He walks under the bridge through
'He walks under the bridge and through it.'

Koopman's analysis of this structure is shown in Figure 6.4. The top projection is PathP—as it must be for any directional PP. This PathP is assumed to take a full C(Place)P complement, which holds the locative PP *onder de brug*. This accounts for the *pre*posed preposition *onder*. The *post*posed element *door* is assumed to be a lexicalisation of the head of PathP. To ensure this appears at the end of the phrase, the whole C(Place)P complement is assumed to raise into the specifier of PathP.[8]

This analysis makes reasonable predictions about how material can be extracted from a circumposed PP. Recall from Section 6.1.1 that the highest specifier in a PP is its 'escape hatch': that a constituent reaching this position can move outside the PP. In particular, it can move to the specifier of a CP position, where it dominates the whole clause containing the PP: this happens in structures representing questions. Koopman's analysis predicts that the regular prepositional phrase (*onder de brug*, in Example 6.38) can be extracted in this way, leaving the postposed P-like element (*door* in Example 6.38) *in situ*. And this is indeed possible, at least for some Dutch speakers.

[8] It's not clear to me why the Path head can be lexicalised in this instance, but not in other cases, such as those discussed in Section 6.1.4.1 where an empty Path head has to be externally licensed. The answer appears to be that *door* ('through') denotes a path 'by itself', while in a PP with an empty Path head (e.g. *in de stoel*, 'in the chair'), there is simply no element explicitly denoting the path.

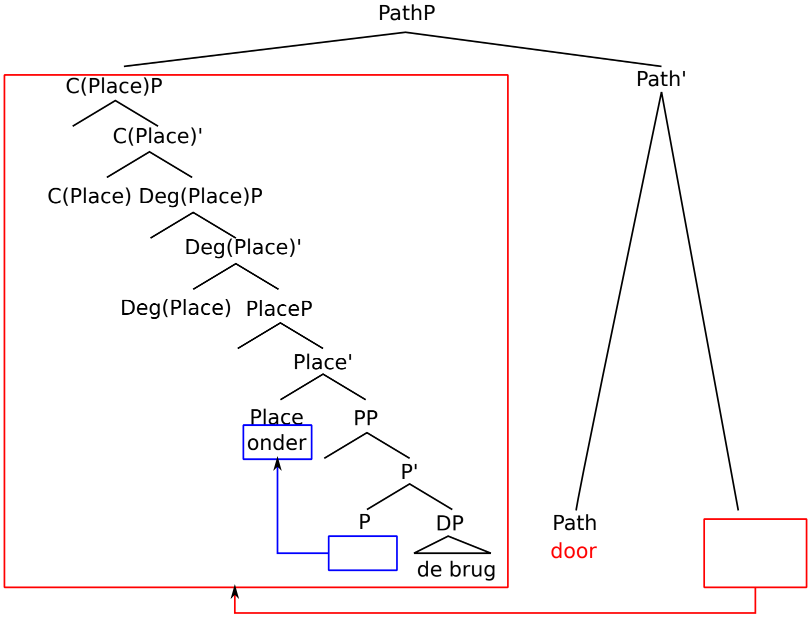


Figure 6.4: Koopman's analysis of a circumposed directional PP *onder the brug door* (*under the bridge through*). The top projection is PathP. PathP takes a full C(Place)P complement, which holds the locative prepositional phrase *onder de brug*. This C(Place)P raises in its entirety to the specifer of PathP. The postposed P-like element *door* is read from the head of PathP.

(6.39) [Onder welke brug] is Jan door gelopen?
[Under which bridge] is Jan through walked
'Which bridge has Jan walked through-and-under?'

However, extraction of the DP *within* a circumposed PP is certainly not allowed—which is predicted from Koopman's analysis, as it involves extraction from within a 'left branch', which is always disallowed (as already discussed in Section 6.1.1).[9]

(6.40) ?[de brug] die Jan onder door is gelopen
[The bridge] that Jan under through has walked

(6.41) *[het huis] dat Jan om heen is gelopen
[The house] that Jan around 'away' has walked

Finally, if we substitute the DP with the R-pronoun *er* and raise it to [Spec,PlaceP] (as shown in Figure 6.1b), and then raise it further to [Spec,C(Place)P] (as shown in

[9] Example 6.41 is more clearcut than Example 6.40; the marginal acceptability of Example 6.40 is due to another possible analysis of *onder door* as a simple postposition. I include it just to maintain the example that runs through this section.

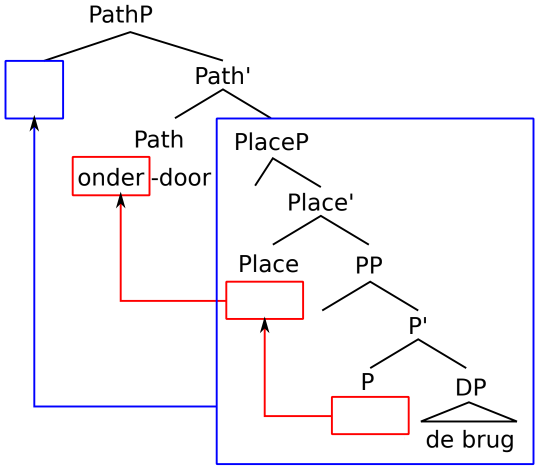


Figure 6.5: den Dikken's analysis of 'amalgamated' postpositions: in this case *de brug onderdoor* (*the bridge under-through*).

Figure 6.2), we can extract this R-pronoun successfully, stranding both prepositions, as shown in Example 6.42.

(6.42) [de brug] waar Jan onder door is gelopen
[The bridge] where Jan under through has walked
'the bridge where Jan walked through-and-under'

A final variant on circumpositional PPs featuring **amalgamated Ps** is illustrated in Example 6.43.

(6.43) Hij loopt de brug onderdoor.
He walks the bridge under-through
'He walks through-and-under the bridge.'

This example comes from den Dikken (2010), who proposes an analysis based on Koopman's model, shown in Figure 6.5. In this structure, PathP takes a PlaceP complement rather than a full C(Place)P. The structure retains the 'lexicalised' Path head *onder*, as in the analysis of a circumpositional PP (see Figure 6.4). However, since PathP directly introduces PlaceP, the P head *onder* can raise up (through Place) to *adjoin* to the Path head, giving the morphologically complex *onderdoor*. As in Figure 6.4, the complement of PathP raises to its specifier position, so that the DP *de brug* appears before the complex preposition.[10] This analysis is quite elegant, as it makes use of the same head-raising operation that sanctions postpositional directional PPs, as discussed in Section 6.1.4.2, coupled with the lexicalised Path head posited in circumpositional directional PPs, discussed in Section 6.1.4.3.[11]

[10] It's crucial that this raising-to-specifier happens *after* the head-raising, of course, though den Dikken does not discuss this.

[11] I guess to analyse the English *in-to the chair*, we could likewise have P raising (through Place) to adjoin to Path. P would hold *in*, which denotes a location, and Path would hold *to*, which denotes a path,

6.1.5 den Dikken's extension of Koopman's analysis

6.1.5.1 Functional projections above PathP: C(Path)P and Deg(Path)P

den Dikken (2010) argues there are functional projections above PathP, that have a similar role to those that introduce PlaceP. He gives several arguments for these projections; I will mention two of the key arguments.

Firstly, den Dikken points out a systematic ambiguity in Dutch degree modifiers: they can modify a location, as we have already seen, but they can also modify a *path*. Consider this example:

(6.44) Het vliegtuig vloog tien meter boven het strand (langs).
The aeroplane flew ten metre above the beach (along)

This can mean the plane flew at a *location* that is situated ten metres above the beach: the modification here is similar to that in Example 6.20 (*[Tien meter er naast] heeft Jan gezeten*). But it can also mean the plane flew along a *path* that was ten metres *in length*. (This is true even if the optional P *langs* is omitted.) This suggests that PathP can be modified.

Secondly, the location and path introduced by a PP can be *simultaneously* modified, as the following example shows. (The PP is fronted, to clearly delineate it.)

(6.45) [Tien meter lang tien meter hoog] vloog het vliegtuig boven het strand (langs).
Ten metre long ten metre high flew the aeroplane above the beach (along)

Moreover, the path modifier has to occur before the place modifier:

(6.46) *[Tien meter hoog tien meter lang] vloog het vliegtuig boven het strand (langs).
Ten metre high ten metre long flew the aeroplane above the beach (along)

This suggests there are separate structural positions for path and place modifiers—and that the position for path modifiers is higher than that for place modifiers.

The most natural proposal is that PathP is dominated by a projection holding path modifiers (**Deg(Path)P**), just as PlaceP is dominated by a projection holding place modifiers (Deg(Place)P). This makes a parisimonious proposal about the general syntax of modifiers, and at the same time, since PathP is above PlaceP, it explains why Path modifiers have to precede Place modifiers, as shown in Examples 6.45 and 6.46. It also neatly explains a fact about modifiers of *amalgamated* Ps of the kind described in Section 6.1.4.3. Such modifiers unambiguously modify path rather than place. Consider the following example:

(6.47) De jongen rende tien meter de luifel onderdoor.
The boy ran ten metre the awning under-through

and these two elements can come together, to create an amalgamated P.

This sentence can only mean that *the length of the boy's path* under the awning was ten metres—not that the location of this path was at a point ten metres below the awning. den Dikken explains this with reference to the analysis of amalgamated postpositions in Figure 6.5. The lower P (*onder*) can only raise to adjoin to the higher P (*door*) if Path takes PlaceP directly as its complement, without PlaceP's functional projections: in this case, the only location where a modifying functional projection can appear is above PathP.

den Dikken also presents good evidence for a **C(Path)P** that sits above Deg(Path)P, analogous to the C(Place)P that sits above Deg(Place)P. This evidence comes from the possible positions for R-pronouns in relation to Path modifiers, illustrated below:

(6.48) Het vliegtuig vloog (er) tien meter lang (er) tien meter hoog (er) boven langs.
The plane flew (there) ten metre long (there) ten metre high (there) above along

The r-pronoun *er* can appear in three possible locations here. The second and third positions are the positions discussed earlier, illustrated in Figure 6.2: [Spec,C(Place)P] and [Spec,PathP]. The first position, underlined, has to be above Deg(Path)P: den Dikken argues, reasonably, that it is [Spec,C(Path)P].

den Dikken envisages one final XP within his prepositional phrase. Recall that the functional projections C(Place)P, Deg(Place)P and PlaceP provide a left periphery for a *lexical* projection, PP, headed by a locative P. By analogy, den Dikken suggests that C(Path)P and Deg(Path)P and PathP should also provide a left periphery for a lexical projection—another PP, but this time, headed by a P denoting a path. This higher path-denoting P takes as its complement C(Place)P—or in postposed or amalgamated PPs, PlaceP directly. The complete structure for a prepositional phrase, with all possible constitutents included, is shown in Figure 6.6.

In Koopman's analysis of circumpositional PPs (see Section 6.1.4.3), one idea was that the postposed 'P-like element' *door* is 'a lexicalisation of the head of PathP'. Now that PathP introduces a *bona fide* lexical projection (PP, headed by P_{dir}), I presume we can situate this lexical material at P_{dir}: then it's unequivocally a preposition. This idea is not explicit in den Dikken, but is a natural modification when P_{dir} is added as the complement of PathP.

6.1.5.2 The semantics of functional projections within the PP

Having motivated a large set of functional projections syntactically, we are then drawn to ask what the *semantic* contributions of these projections are. den Dikken's proposal here relates to a rather sketchily introduced general account of the extended projections dominating verbs, nouns and prepositions. I will begin by outlining the generalisation.

It is relatively uncontroversial that the lexical projection introducing a verb, VP, is dominated successively by functional projections for Aspect, Tense and Force. Features in the Aspect head encode the distinction between bounded events and unbounded events (e.g. the distinction between *John walked* and *John hiccupped*) (reference). Features in the Tense head encode deictic references to the present, past or future (Partee, 1984). Features in the Force head encode the distinction between assertions, questions and imperatives

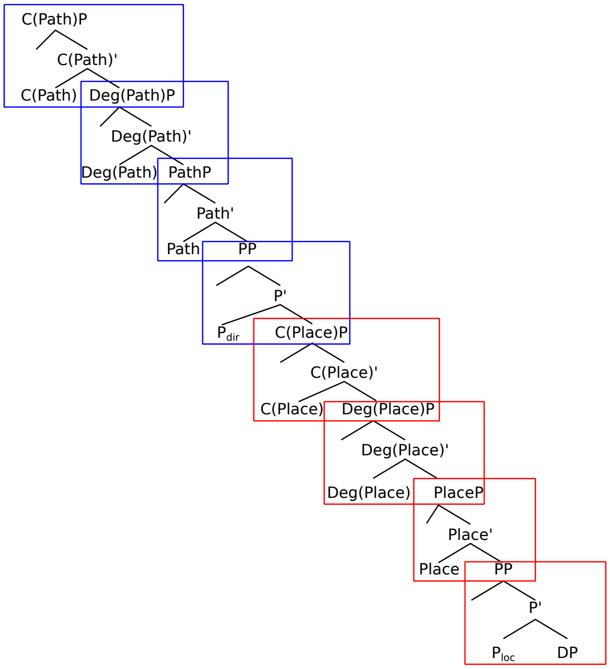


Figure 6.6: The complete X-bar structure of a directional prepositional phrase, in den Dikken's analysis. The extended projection of the locative P (P_{loc}) is highlighted in red; that of the directional P (P_{dir}) is highlighted in blue. Note the parallelism between the two extended projections: they both have the structure CP→DegP→XP→PP.

(reference). den Dikken argues that an NP is dominated by a similar trio of functional projections. NumP (Ritter, 1991) encodes the distinction between 'mass' and 'count' referents; PersonP (citation) encodes deictic references to the speaker, hearer, and third-party individuals; DP encodes definiteness, which is comparable to illocutionary force. Extending this pattern, den Dikken argues that PathP and PlaceP are aspectual projections, encoding the distinction between bounded and unbounded paths and places, Deg(Path)P and Deg(Place)P are deictic projections, encoding the distinction between places or paths near-to or far-from the speaker or hearer, and C(Path)P and C(Place)P encode something in the spatial domain analogous to illocutionary force (for verbs) and definiteness (for nouns). den Dikken does not offer an account of the semantics of C(Path)P and C(Place)P, which is still a matter for research, but he does give interesting accounts for PathP/PlaceP and Deg(Path)P/Deg(Place)P, which I will summarise below.

PathP and PlaceP: bounded and unbounded paths and places den Dikken suggests that features at PathP encode the difference between bounded and unbounded paths. It's clear that this is a point of difference between path-denoting prepositions: for instance, *to* X denotes a bounded path, with location X as its fixed endpoint, while *around* X encodes an unbounded path, within the environment X, or using X as a landmark. I presume the bounded/unbounded nature of these paths is encoded at the Path head, while some more open-class aspect of their semantics is encoded at the lexical P_{dir} head that Path introduces.[12]

den Dikken proposes that PlaceP encodes a similar bounded/unbounded distinction, this time between places rather than paths. He refers to Tortora's (2006) discussion of Italian and Spanish, where a distinction along these lines is made explicitly. Tortora shows various contrasts, including this one:

(6.49) Vai a giocare [dietro a quell'albero].
Go to play behind 'a' that tree
'Go and play behind that tree.'

(6.50) *Vai a giocare [dietro quell'albero].
Go to play behind that tree

The PP *dietro a [DP]* contains a regular locative P *dietro*, and a particle *a*. Tortora argues the function of *a* is to signal an unbounded location. The role of the unbounded location in Example 6.49 is to identify an unbounded location in which the hearer is instructed to play. The assumption is that playing involves some degree of movement, so it's not possible to play at a *specific* location. What is being located 'behind the tree' is a *set* of locations—or perhaps, in my terms, the location of a *sub-environment* near the tree. Example 6.50 is ill-formed, because the PP dietro all'albero, having no *a* particle, indicates just a single location.

Tortora makes use of den Dikken's proposal that the head of PlaceP encodes the distinction between bounded and unbounded locations: she argues the particle *a* appears at this head position. Deriving the surface PP *dietro a [DP]* is a two-stage process, similar to that envisaged by den Dikken for amalgamated postpositions (see Figure 6.5). First, the DP moves out of the PP to a higher Spec position; next, the PP (holding just the P *dietro*) moves to the Spec of PathP, so it appears before *a*. The analysis is shown in Figure 6.7. Note that this analysis requires an extra FP in between PlaceP and PP, so that the DP can raise to a specifier position that's below PlaceP, but still outside PP. In fact in Koopman's analysis there is an FP in this position, an agreement (Agr) projection, which I omitted in my presentation. den Dikken does not see the need for an Agr projection here, but whatever its role, it is needed to make Tortora's analysis work.

Deg(Path)P and Deg(Place)P: speaker/hearer-oriented representations of path and place Most languages have ways of encoding path and place 'deictically', in relation to the location of the speaker: in English, for instance, this is done with the deictic PPs *here* and *there*. However, these expressions can modify *both* path and location.[13] den Dikken cites German as a language that has explicit 'deictic' (i.e. speaker-oriented) modification

[12] The lexical P_{dir} in any case *carries* Path features, and raises to Path to 'check' these, so both aspects of its meaning are encoded at both heads.

[13] den Dikken notes that the archaic *hither* and *thither* are unambiguously path modifiers. Presumably these would appear at the head of Deg(Path)P, though he does not say so explicitly.

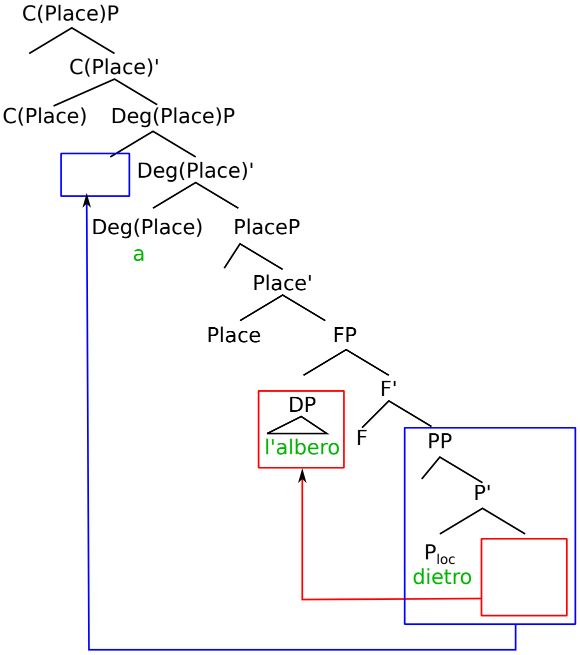


Figure 6.7: Tortora's analysis of the 'unbounded' locative PP *dietro all'albero*.

of path. In particular, the particles *her* and *hin* denote paths towards and away from the speaker: unlike the English *here*/*there*, they cannot be used to describe static locations. den Dikken (2003) gives a detailed analysis of the structure of German PPs, using the extended projections introduced here (summarised in Figure 6.6). I won't describe this at any length, but the proposal that *her* and *hin* appear at the head of Deg(Path)P fits well with other elements of this model. In particular, it gives an account of German circumpositional PPs, featuring postposed constructions such as *hinauf* and *herunter*, illustrated below:

(6.51) Die Schnecke krocht [auf das Dach {hin/her}{auf/über/unter}]
The snail crept [on the roof {towards-me/away-from-me}{on/over/under}]
'The snail crept {hither/thither}{up/across/down} onto the roof'

den Dikken argues the regular path prepositions *auf*, *über* and *unter* sit at the head of PathP, where they introduce the locative PP *auf das Dach*.[14] He argues the speaker-relative particles *her-* and *hin-* sit at the head of Deg(Path)P, just above PathP. To derive surface structure, the regular path preposition raises to adjoin to the speaker-relative particle, creating an amalgamated expression, and the locative PP raises as a whole to [Spec,Deg(Path)P], so it is pronounced before the amalgamated expression, as shown in Figure 6.8. Note in this analysis, there is some ambiguity in the word *auf*: the word as it appears at the head of P_{loc} (in *auf das Dach*) means 'on', while the word as it appears at the head of Path means 'up'.

[14] den Dikken doesn't discuss how much of the extended projection of this PP is introduced: I'll assume none, to keep things simple.

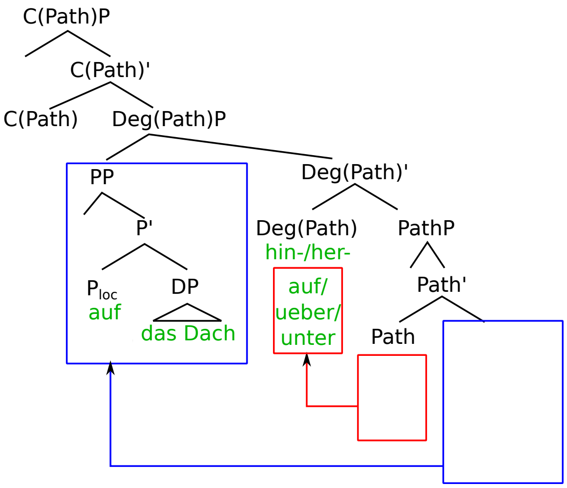


Figure 6.8: den Dikken's analysis of *auf das Dach* {*hin/her*}{*auf/über/unter*}

den Dikken argues that in Hungarian, separate PP-internal particles systematically mark path and place deictically, in relation to the location of the speaker: *itt* and *ott* unambiguously denote static locations near and far from the speaker.[15]

A final point: recall that DegP was originally introduced (by Koopman) as the projection hosting degree expressions in its specifier. If DegP also hosts deictic references to the speaker, we have some work to do to explain what these elements have in common.

[15] I need to look into this in more detail: den Dikken just touches on it.

Chapter 7

A SM interpretation of PPs

7.1 Towards a SM interpretation of the Koopman/den Dikken model of PPs

The Koopman/den Dikken (K/D) account of the structure of PPs refers to several standard ideas in syntactic models. Some of these relate to representations. In Koopman's account, for instance, the head of PlaceP hosts a 'feature' called Place. Others relate to movement operations, and the circumstances that compel them. For instance, the reason why PPs and R-pronouns can raise to [Spec,PlaceP] is that they are 'attracted' to this position. (R-pronouns are attracted there because they 'have' this feature—they are 'morphologically specified for Place'. I presume that PPs also have it in some sense, though it can't be a morphological one.) The reason why DPs cannot raise by themselves to [Spec,PlaceP] is that they are *not* 'attracted to' [Spec,PlaceP]—i.e., I presume, they *don't* have the feature Place. As in all syntactic models, these bits of theoretical machinery are purely formal; there's no attempt to link them to anything outside the linguistic system.

7.1.1 An initial idea about the SM interpretation of the extended projections C(Path)P and C(Place)P

Consider first the extended projection of directional P (e.g. *to*). One nice idea is that the high Specifier position of this projection, C(Path)P, denotes a *goal location*, whose activation triggers a locomotion action. The complement of directional P could then denote an *actual location*, at the point when this is *identical with* the goal location. At this point, some learning happens: the actual location is 'copied to' the goal location, and we can train the function that generates the goal location from some distal perceptual representation.

This can perhaps provide a nice account of the kind of raising found in circumposed directional PPs like *onder de brug door* (see Section 6.1.4.3, Figure 6.4). Here, the full extended projection of a locative P, C(Place)P, raises into the Specifier of a higher XP. In Figure 6.4, the higher XP is PathP, which is the highest projection in Koopman's analysis, but within den Dikken's model, we can imagine the C(Place)P raising right to the Specifier of C(Path)P. On this model, the two linked positions conceivably correspond to two references to a given location: one as navigation goal, and one as current state.

I think here we have to envisage that the extended projection of a locative P represents a distinct planning medium from the extended projection of a directional P. We can conceive of the extended projection of a locative P as representing a (rehearsed) locomotion action, and the extended projection of a locative P as representing a (rehearsed) environment/place-update action. When rehearsing a locomotion action, we have two opportunities to rehearse an environment/place-update action, which is associated with a location: first when this location is activated as a goal location, and later, when it's activated as the current location.

7.1.2 Thinking about the SM denotations of XPs within C(Place)P: some background ideas

A locative PP identifies the environment and/or location of a 'subject', relative to a specified individual. The subject is supplied by syntactic context: in the above examples, where the PP is the predicate of a sentence, the individual is the grammatical subject of the sentence (the cup). The specified individual is given by the DP complement of the PP. The P and DP can combine in different ways to identify the environment and/or orientation of the subject. Here are some examples:

(7.1) The cup is [on the table].

(7.2) The cup is [by the coffee pot].

(7.3) The cup is [20cm beyond the coffee pot.]

In Example 7.1, the complement DP identifies the subject's environment, and nothing is said about the subject's location within this environment. In Example 7.2, the complement DP identifies another individual in the environment, and the P indicates a particular spatial relationship (proximity) between this individual's location and the location of the subject. Example 7.3 is similar, except the spatial relationship denoted by P ('beyond') indicates only a direction: a modifier (*20cm*) indicates the distance, to select a specific location. I'll consider these examples one by one.

7.1.2.1 *The cup is on the table*

This sentence answers the question 'Where is the cup?'. The speaker is someone who has already incorporated the cup into his spatial representation scheme. If the sentence was an existential, *There is a cup on the table*, it would report the process whereby an observer does this incorporation. I'm hoping *The cup is on the table* reports a simpler process.

I think the process has to involve something like this. The observer has to identify the environment that the cup is in (in this case, the table), and the spatial relation between this environment and the cup (in this case, 'on'). To establish 'on', the observer must find a surface on the cup (probably its base) that coincides with a surface in the environment (probably the tabletop). Actually, he also needs to establish more detail, that doesn't get conveyed in the surface form of the PP—but might be present at LF. He needs to establish (i) the *place* on the tabletop where the cup is; and (ii) the orientation of the cup on the tabletop (i.e. the angle between the coordinate system of the table and the cup). But this information is underspecified by the preposition *on*.

There are a few separable operations here. (i) Identifying the environment (the table). (ii) Identifying a surface in that environment (the tabletop). (iii) Identifying a place in this surface. (iv) Identifying a surface in the cup. (v) Identifying the congruence of two surfaces (to get 'on'). These operations could conceivably be denoted by separate XPs in a locative PP.

All of these operations relate to the table, but also to the cup. The table is selected because it's where the cup is. Ditto the tabletop, and the place on the tabletop. The base of the cup is obviously on the cup. And the 'on' relation is also obviously a relation between the table and cup. One idea is that *each* XP denotes an operation relating two things. Everything should come together at the final rightmost XP, which denotes a learning opportunity.

Say we begin with spatial attention on the cup. Say the first operation in the locative PP is selecting the table. This operation must be *referred to the focus of spatial attention*. How does this happen? One idea is that the operations in a locative PP make reference to an *egocentric* (i.e. *perceptual*) representation of the cup's location. Say the relevant perceptual modality is the one that maps retinal representations onto *motor states*, understood in a broad sense as patterns of activity in the SOMs that hold 'hidden states' of motor routines (both for navigation and for movement of effectors, especially hands). Say in the initial context, the observer maps every point on the retina that projects something 'out of reach' to the SOM responsible for navigation in the current environment, and every place that's in reach for effector E_i to the SOM responsible for moving E_i. In this context, when the observer attends to the cup, and gets a retinal representation of the cup's location (and shape), these can be mapped to *motor states* representing location. In the case of the cup on the table, the retinal region associated with the cup is associated with a location in the 'navigation' cognitive map. It happens that there is a known object at that location: the table. This becomes active when the location is activated. However, the region associated with the table is larger than the region associated with the cup. So we repeat the process. We first establish the table as an environment: specifically, a collection of surfaces. I assume these surfaces are specified in motor coordinates. I'm not sure exactly how this happens, but the most obvious idea is that the agent imagines he is 'at' the table (specifically, that he has walked *to the table* via the most obvious path). At this point, the surfaces of the table can be specified in motor coordinates of the agent's effectors. Independently of this, each surface of the table is associated with its own retinal region, and we see which of these is the one associated with the cup. We choose the region associated with the tabletop. This has the effect of selecting one particular effector-based motor system: the one used to navigate a horizontal tabletop-like surface. But again, the retinal region associated with this motor system is larger than that associated with the cup. So we establish the tabletop as an environment, to activate a set of places within the tabletop. The cup corresponds nicely to one of these, so we're almost done. But we have to finish by establishing a particular *relationship* between the tabletop and the cup. In this case, the relation involves a particular surface on the cup, which coincides with the surface of the the tabletop (locally, at the place where the cup is). To detect this coincidence, we need to make use of the motor definition of the tabletop surface. The agent can represent the location/orientation of the tabletop surface at the place where the cup is as a motor state; he can also represent the location/orientation of surfaces in the cup as motor states. So he can discover the correspondence between a surface of the cup and the tabletop.

7.1.3 *My hand is on the table*

[Stepping up one level..]

If the spatial relationships denoted by locative PPs are ultimately defined in motor terms, *My hand is on the table* should be a useful example to consider. Here, the observer is the agent. I've already got a model of stable support of the hand on a surface: see e.g. Section 2.3.3. The main extra information conveyed by *My hand is on the table* relates to the identification of the agent's hand, and the table, as *objects*. The 'subject', *my hand*, is outside the PP. I presume the agent's focal attention is on his hand. Let's say this means he is attending to the tactile sensations on the palm of his hand. (That's something he has to be doing to identify a stable support relationship, I guess, so it's not such a big assumption.) Let's say it also means his visual focal attention is on the region of the retina where the hand's image is projected. (That is, I assume he has learned to locate his own hand using vision.) The operations reported by the PP *on the table* are *referred to* this focus of visual/tactile attention. The first operation is to activate a motor representation of the hand's location. This is a representation of a location in the agent's peripersonal space. In this case, this is simply a representation of the agent's current arm state: this represents a location. We now need to know what is *at* this location. At this point, the tactile system is consulted again, this time to deliver information *about a surface in the world*, rather than about the hand as an object. (The hand is now being used as a sensor, rather than represented as an object.)

What is now at issue is the identity of the object the hand is touching. A process of object recognition has to happen. If the agent were blind, he would feel the object, to build a 3-dimensional shape representation, and then match that to a set of types and tokens stored in LTM. If he can see, he can direct another visual process to a retinal region associated with that associated with the portion of the hand receiving a stable contact stimulus.

What is this 'associated retinal region'? Obviously, the visual system has to *learn* something about the region associated with the surface the agent is touching with his hand. In this case, the learning happens when the agent is *exploring* the desktop surface with his hand. While this happens, the agent can learn a visual representation of the support surface that the hand is navigating through.

Generalising away from this one case, the agent can explore several kinds of environment. I'll consider three.

1. With his hand, he can explore the outside of a solid convex 3D object like a solid cube, moving his hand from one surface to another. During this process, he can learn the visual representation that's associated with the whole cube, which he can use to predict the constraints on his hand's navigation movements.

2. With his hand, he can also explore the inside of a concave 3D object—for instance, a rectangular box. While exploring the inside surface of such a box, he will encounter constraints on his hand's movement. Again, he can learn the visual representation that's associated with this particular set of constraints, and get a representation of the whole box, as an 'environment'.

3. Using whole-body navigation, he can explore his *own* local environment. Again, vision can be useful in identifying constraints on navigation—i.e. the shape of the local environment. (And also possibly its type, or identity as a token.)

What does it mean to learn a visual representation of 'an environment', rather than of an object? I think it just means that you map a visual representation onto the motor representation that supports navigation. That comprises (i) a 'holistic' representation of the current environment; and (ii) a representation of your current location *within* this environment. (That feels significant to me: the DP complement of a PP can either denote the whole environment, or a location within that environment.)

There's focal attention on 'the navigating object' (the hand) when its relationship to the table is determined. One question is: what should the attended region associated with the hand's environment be? I think there are two things we can say. Firstly, it should *include* the region associated with the navigating object. Secondly, it should be *larger* than this region. (If the object is navigating, it's moving around several locations, and so will subtend a larger set of retinal regions.) Modulo that constraint, the retinal region associated with the environment of an attended object can still be a *local* region of the retina. It should just be one that's larger than the local region associated with the navigating object.[1] So when you're looking for an object's local environment, you should be looking within a retinal region including the region that contains the object, but larger than it. How much larger is a key question.

I'll first consider how the agent identifies the visual region containing the environment of a given attended object. (In Section 7.1.3.1.) Then I'll consider how the agent learns to represent the geometric properties of this environment, including its orientation in relation to himself, based on the visual features within this region. (In Section 7.1.3.2.) Finally I'll consider how the agent learns to represent the location of an attended object in a coordinate system centred on this environment, given the visual stimulus it projects, plus a retinal representation of the object's location and orientation. (In Section 7.1.3.3.)

[1] Of course it doesn't *have* to be a local region. In the case where the navigating object is the whole agent, it can be the agent's full peripheral vision. That's even the case where the attended object is large enough that its local environment is the same as the agent's local environment.

7.1.3.1 A function that learns to identify a set of retinal locations associated with a surface currently being explored by the hand

I assume that a key structure in the visual representation of an environment—or to be simple, say a surface—is a clearly-bounded *retinal region* that the surface projects onto. I'll call this region an **environment-encoding region**. I assume a special visual function that delivers this region. The key idea about this function is that it delivers a *retinal region*, rather than a semantic representation. In this sense, it's like a saliency function. I propose that the environment represented as a contiguous set of retinal regions provides one representation of the environment 'as an environment': specifically, a set of places. However, if this region can be isolated, it can also function as a mask over the retina. I suggest this mask has two functions. Firstly, it allows geometrical properties of its boundaries to be computed (as discussed in Section **??**). Secondly, it can provide a very accurate gate on the input to the semantic object classifier in inferotemporal cortex (discussed elsewhere).

I'll first consider the kind of network that will deliver the function: I'll propose that it's a spreading activation network. Then I'll consider how motor states relating to navigation are encoded within the region.

Defining the environment-encoding region using spreading activation One idea is that the function implements a 'spreading activation' within the retina, starting from the attended retinal location, which projects the current environment, and spreading in parallel to adjacent retinal locations if there are cues the environment extends in this direction. There are several models that work this way, exploiting lateral connections between cells with neighbouring receptive fields; see e.g. Jehee *et al.* (2007) for a network model. There's some debate about how much of this circuitry is hardwired and how much needs to be learned through experience (see e.g. Lee *et al.*, **?** for evidence they are partly learned); my idea is that the learning happens, at least in part, when the hand is travelling around the environment.

For concreteness' sake, say the function takes the current retinal location $Current$, and (in parallel) each adjacent retinal location Adj_θ, at retinal discrepancy θ (an angle). Specifically, it's taking the *simple visual features* present at each location: $f(Current)$, and $f(Adj_\theta)$ (along with θ itself). The function *Spread* returns a Boolean: *true* if spreading can occur; *false* otherwise.

$$Spread(f(Current), f(Adj_\theta), \theta) : Boolean$$

What are the simple visual features involved? I propose they include: (i) distance, computed by stereopsis and similar methods; (ii) surface normals, likewise computed by stereopsis and similar methods; (ii) linear features (including curvilinear features). The kind of things the function might learn are illustrated below:

- If the Current retinal location has surface orientation O, and the Adjacent location above it (at $\theta = 0$) contains a horizontal linear feature (at the bottom), coincident with a distance discontinuity, then return *false*.

- If the Current retinal location contains a horizontal linear feature (at the top), and the Adjacent location to the right (at $\theta = 90^o$) contains a collinear horizontal feature (also at the top) then return *true*.

One nice aspect of this spreading-activation definition of the environment-encoding region is that there's a natural way to train the *Spread* function *as the hand is travelling through the environment*. Whenever the hand moves from *Current* to Adj_θ, this provides a single training instance for the *Spread* function.

Motor representations within an environment-encoding region I propose that the retinal regions inside an environment-encoding region convey two things to the agent. Firstly, they convey a set of motor states: the visual features in each region should map onto a corresponding surface-based location (expressed within the recurrent SOM representing surface-based hand locations).[2] I think these motor states are generated serially, rather than in parallel, by a function that takes a *selected* region and returns a pattern in the recurrent SOM encoding a place on the tabletop. (The visual features of each region can certainly be generated in parallel, though.) Secondly, again for a selected retinal region, we can envisage a function that identifies the *neighbouring* regions, and for each (if appropriate) generates a *motor command* that will get the hand to the place on the surface projecting to that region. These give the agent a set of local navigation affordances for the hand. (I assume these movement options really are generated in parallel, so that the best one can be directly selected.)

In summary: when you're looking at a tabletop, and generating an environment-encoding retinal region corresponding to the tabletop you can either discretely jump from one point in this region to another, in each case activating the orientation of the surface at that point (expressed as a hand position/orientation), or you can execute a 'visual routine' starting at one of these points, attending successively to the points in the region that would be occupied by a hand travelling around the surface.

A surface like a tabletop can also contain obstacles to navigation, that constrain the movement options at particular points. (The recurrent SOM should learn about such obstacles.) The key retinal region for an obstacle is the one associated with its *base*—that is, the surface of the obstacle that's resting on the tabletop. The retinal regions associated with other parts of the obstacle may simply be *occluding* the surface. I'll talk more about occlusion below.

Occlusion If my view of the tabletop is occluded, for instance by an object sitting on it, or by something crossing in front of it, the tabletop should still be recognised as a single surface. This is called the Gestalt principle of continuity. How does this work?

[2] It should perhaps *also* be expressed as an absolute motor state of the arm/hand, in which the hand has stable contact on the surface projected by that region: that is, in an *agent-centred* coordinate system. In that case, the agent systematically represents points on the surface in two coordinate systems: firstly an object-centred coordinate system, centred on the tabletop; secondly, an agent-based coordinate system, centred on the agent.

If we use spreading activation to define the boundaries of the tabletop (see above), one possibility is that activation is 'tentatively' extrapolated *beyond* an apparent boundary. If it 'meets' activation tentatively extrapolated in the other direction, this counts as evidence for a continuous surface occluded by something (rather than two separate surfaces).

An interesting case of occlusion is for a curved surface: here, the surface can occlude itself. In this case, a given retinal region should be mapped to two different, discontinuous, motor states. At least if the surface has uniform curvature, the agent can have a good idea about the location/orientation of the hidden part of the surface, from visual properties of the visible part (e.g. shape-from-shading, the linear shape of the visual object boundary).

7.1.3.2 Representing the geometrical properties of the surface occupying an environment-encoding region

Assume the agent has activated an environment-encoding region representing a surface (a tabletop in our example), and is therefore able to simulate movements getting discretely to each point on the surface, and, for any given point, to simulate a set of possible trajectories within the surface from that point. The agent still needs to be able to use vision to represent the location of his hand on the surface, in a coordinate system centred on the surface. This is hard, because both the environment-encoding region and the region containing his hand are given in retinotopic coordinates. The agent has to learn about many possible orientations of the surface separately.

I assume this happens in two stages. The first stage is to encode the visual properties of, or within, the environment-encoding region, that indicate the orientation and intrinsic geometry of the surface projecting onto it. I'll call these properties the environment-encoding region's **geometry properties** I'll talk about the first stage in this section. The second stage is to pass these properties, together with the retinotopic position and orientation of the attended object, into a function that delivers a surface-centred representation of this position and orientation. I'll talk about the second stage in Section 7.1.3.3.

The first stage is to compute the geometry properties of the environment-encoding region. I will assume these comprise two kinds of properties. One kind are the linear properties of the *boundaries* of the environment-encoding region. (Here talk about Mary Peterson's stuff.)

The other type of geometry properties are the features of the regions inside the environment-encoding region that inform about the orientation and curvature of the surface it represents.

The geometry properties of the environment-encoding region can be used to identify the region, for instance as a shape (e.g. a square, a rectangle), but also as a token environment. I'll consider token environment recognition first. There's a big literature on 'how environments are recognised'. Obviously there's more to it than recognising a shape/geometry, but shape/geometry is at least one feature. (You should also mention the other things that are relevant—in particular textures, that identify what the environment is made of, and landmarks.) We can envisage an **visual environment recognition function**, that takes as input these geometry properties, and delivers as output an *LTM environment*. Recall that an LTM environment is a unit that delivers a static bias on the dynamics of the recurrent SOM that encodes the agent's place in the environment (I'll call that the **places SOM**). If the agent's visual attention is maintained on a given surface (i.e. if the agent is 'tracking' the surface)[3] while he changes his orientation with respect to it, he can generate good training date for the visual environment recognition function.

7.1.3.3 Representing the hand's location within an environment-encoding region

Having computed an environment-encoding region's geometry properties, these can be passed, along with retinal information about the hand (including both location and orientation), to a function that computes the hand's location in a coordinate system centred on the surface. I'll call this function the **visual hand location recognition function**. The key thing about this function is that it's trained to produce the activity in the agent's places SOM. If the agent's visual attention is maintained on the surface, *and the agent's hand is kept at a fixed position on the surface*, there will be a static pattern of activity in the places SOM, while the surface's geometry properties and the hand's retinal location and orientation both change—so the function can learn all the necessary invariances.

7.1.3.4 Representing movements of the hand outside the navigation surface

Say the surface being explored by the agent's hand contains an obstacle: for instance, a heavy book in the middle of a desk. In the case of whole-agent navigation, there are trajectories that take the agent *round* such obstacles—and there are similar trajectories in the case of hand navigation. However, for hand navigation there is also another option: to go *over* the obstacle. This route involves the hand temporarily *leaving* the navigation surface, and returning to it at a different point, via a trajectory through a point above the obstacle. My question here is: how can the agent represent this trajectory? Or more generally, any trajectories that pass *over* the surface, rather than along it? (Note the trajectory passes 'over' the obstacle object as well as over the surface.)

Recall from Section 7.1.3.1 that the environment-encoding region associated with a surface comprises a set of retinal locations, each of which is associated with an agent-centred goal motor state, which will put the agent's hand into contact with the surface at the point that projects to this location. I suggest that the key representations for controlling movements that leave the surface are these goal motor states. Say the agent's hand is currently at place P_1 on the surface, and wants to get to place P_2. One way to do this is to navigate *on the surface* to P_2. Another way is for the agent to lift his hand off the surface, and simply activate the goal motor state associated with P_2, to generate a 'reach' movement culminating in the hand being at P_2.

The via-point for this reach movement is equally important: it has to be such that the hand avoids the obstacle (the book). I suggest the hand first has to travel to 'a place above the book', and then back down to P_2. I think the book is defined as a place, rather

[3] Tracking a surface involves tracking a region, across retinal translations but also changes in shape/size. I should come up with a model of how this happens.

than as an environment. But this place has to have a 3D character: it has to be a volume (including height), rather than just an area on the navigation surface. I'll have an interlude to discuss 3D environments before I consider how places are represented in 3D.

7.1.4 3D environments

How does the agent represent a whole object as a navigation environment? I suggest

7.1.5 Some initial ideas about SM denotations of XPs within C(Place)P

I'll begin by thinking in more detail about the C(Place)P, Deg(Place)P, PlaceP sequence. My hypothesis is that this denotes an environment/place-update action. Consider a full PP related to the ones discussed in Section 6.1.1, and a similar English PP:

(7.4) [Tien meter naast de deur] heeft Jan gezeten.
Ten meters next-to the door has Jan sat

(7.5) John's house is [ten metres past my house].

In my hypothesis, *Tien meter naast de deur* reports an update of the spatial representation system: specifically, the selection of a new 'current LTM environment', and then a particular place *within* this new environment. I don't think the environment is identified with a DP: the environment is certainly not the door. The door is a landmark in the environment. In the case of *ten metres past my house*, the environment is presumably *the street* on which both John's house and my house both sit.) I assume there's a contextually-accessible environment which supplies a sensible denotation for *naast* (or *past*), and *Tien meter naast de deur* (or *ten metres past my house*) reports the establishment of this environment, as well as the activation of a specific place (and possibly orientation) within this environment. Now the question is: what operations within this process might C(Place)P, Deg(Place)P and PlaceP denote?

A possible SM interpretation of C(Place)P I argued above that C(Path)P denoted the goal location for a locomotion action. I don't think the C(Place)P can denote a goal—at least not for a bare locative PP. (Maybe it would if the locative PP occurred within a directional PP, but not for a simple locative.) So there has to be some other denotation for C(Place)P. Perhaps it could straightforwardly denote the establishment of a new environment: in the case of Example 7.5, for instance, the environment in which John's house (and my house) are situated.

On that hypothesis, what might the head of C(Place)P denote? It could perhaps denote the newly activated LTM environment itself. Perhaps there are no interfaces from LTM environments to phonology, so environments can't be directly described—only their associated objects. (In any case, here the environment-qua-object is also implicit: all we have to infer it is a landmark object *in* the environment.)

What might the Specifier of C(Place)P denote? Note that this is one of the positions where the R-pronoun *er* can appear, in the K/D model. This is the clitic-like PP pronoun that means something like *there*. It also might be a position to which a PP with a full DP can raise in German and Italian (see Section 7.3.2 below), though apparently not in Dutch. On my general interpretation of specifiers, in C(Place)P denotes the activation of a new LTM environment, [Spec, C(Place)P] denotes the 'reafferent sensory consequence' of this new activation. What might this be? It could perhaps be a location *within* this new environment. Perhaps this location could be defined retinotopically, rather than allocentrically, and a later location representation is given allocentrically, and this is the basis for the multiple possible positions of PP.

A possible SM interpretation of Deg(Place)P Let's say the 'start context' for Deg(Place)P is a newly active LTM environment and a newly active location within this environment. I assume locations (places) also lack a direct interface with phonology, so they can't be verbalised in any meaningful way. But one thing a speaker can say about a location is whether it is near or distant from him/herself. den Dikken argued that some languages (e.g. Hungarian) verbalise this on the Deg(Place) head. This head could express an operation that compared the selected location in the new environment to the speaker's current location. (This even makes sense if the new environment doesn't contain the speaker: I can talk about two points on a desk as 'here' and 'there', even if I'm not on the desk myself. But it doesn't work for all environments: if I'm describing two places in my kitchen, and I'm far from my kitchen, the terms 'here' and 'there' aren't applicable.)

Another way of identifying a place P is to specify the trajectory between a landmark with a known place and P. This is particularly useful when the preposition just denotes a direction. In a case like *ten metres past my house*, *past* supplies the direction of the trajectory, and *ten metres* supplies its length.

In our model of locomotion, I'm sure we can measure distances in units: for instance, there's a natural unit of a single iteration, so the instruction 'go forward for n units' has a very natural interpretation. But I'm not sure why this representation occurs as 'high' as Deg(Place)P. If the 'start context' of an XP involves an active place and an active environment, and its head denotes activation of a reference place, I can see how a degree expression like *ten metres* would show up in the specifier of an XP: the vector would be activated as a side-effect. But the XP whose head activates a reference place is surely lower down than Deg(Place)P in the syntax. (...)

7.1.6 SM interpretations of head movement and XP movement: recap

I'll start off with by summarising my general SM interpretations of movement. There are two of these: one for head movement, one for XP movement.

7.1.6.1 Head movement

For me, a right-branching sequence of XPs denotes a replayed sequence of SM operations. Heads of XPs can move because they denote SM operations as they are represented in the planning medium, where they are active in parallel throughout the replayed sequence.

7.1.6.2 XP movement

XP movement (e.g. the movement of a terminal XP to a higher Spec position) denotes something different. In this case, the XP can appear at two syntactic positions. The higher (Spec) position represents the reafferent sensory consequence of a SM operation; the lower (Complement) position represents the SM context brought about by a (later) SM operation. In my SM model, these two positions represent 'the same thing' in two different modalities, and there are mechanisms that learn associations between them. After this learning, a single multimodal representation is activated twice during rehearsal, once early in rehearsal, and once at the end of rehearsal, in the terminal context. In my sentence generation model, when this representation becomes active, it activates a representation in a *different* planning medium (the WM individuals medium), and thereby creates an opportunity to switch to a different mode of control, and rehearse the activated plan in this medium. Since the representation is activated twice, it provides two opportunities to switch to this secondary mode of control: an infant learning a language has to learn which opportunity is conventionally taken for that language.

In the special case where the XP is an R-pronoun, if this XP is raised to the *highest* Spec position within the pp, it can raise *outside* pp to a position within the clause, associated either with VP (the verb cluster) or CP (in questions).

7.2 Ideas about head movement for PPs

A key idea in the K/D model is that the raising of P is subject to certain limits. It can always raise to the head of PlaceP. Whether it can raise higher depends on what XP introduces PlaceP. If PlaceP is introduced by Deg(Place)P, its raising is blocked. But if it is introduced by PathP, it can raise to the head of PathP.

7.3 Ideas about XP movement for PPs

In the K/D model of PP structure, there are four types of XP movement: the moving item can be DP, PP, PlaceP or C(Place)P. In each case, the XP is a *terminal* one within the chain of XPs, and the whole XP moves; in each case, it moves to a higher Specifier position. It doesn't have to be the nearest Specifier position.

7.3.1 Raising of PP to [Spec,PlaceP]

In locative PPs, PP can raise to [Spec,PlaceP] (see e.g. Section 6.1.1, Figure 6.1(a) and (c)).[4] Assume, as suggested above, that C(Place)P, Deg(Place)P and PlaceP represent a rehearsed environment/place-update action within a planning medium holding planned environment/place-update actions. To interpret PP-raising *within* this structure, I must assume that the PP unit expresses a rehearsed routine within yet another planning medium—and that there are two opportunities to access this routine: one at an early point during rehearsal (corresponding to [Spec,PlaceP]); another at the end of rehearsal (corresponding to [Comp,PlaceP]). The question is now: what planned routine can PP represent, that's distinct from the planned environment/place-update action represented by C(Place)P, Deg(Place)P and PlaceP? And what's the commonality between [Spec,PlaceP] and [Comp,PlaceP], which means that a PP can appear in either position? (And why can a PP with a full DP raise *no further* than [Spec,PlaceP], while an R-pronoun PP can raise to several higher Spec positions?)

7.3.2 Raising of a (full) PP to [Spec,Deg(Place)P]

This happens in Tortora's analysis of Italian *dietro all'albero* (see) and in den Dikken's analysis of German *auf das Dach hinunter* and friends (see).

7.3.3 Raising of PP to [Spec,Deg(Path)P]

7.3.4 Raising of PlaceP to [Spec,PathP]

7.3.5 Raising of DP to [Spec,FP]

This happens in Tortora's analysis of Italian *dietro all'albero* (see Section 6.1.5.2; Figure 6.7).

7.3.6 Raising of DP to [Spec,PathP]

7.4 Old below

7.5 Basic idea

Consider the following LF analysis for *John climbed onto the table.* (Assume John is the observer.)

(7.6) John [$_{VP}$ walked [$_{PathP}$ to [$_{DP}$ the table]]]

[4] The unit that raises might actually be the FP that dominates PP, as hypothesised by Tortora (see Section 6.1.5.2, Figure 6.7); but I'll refer to PP in my discussion here.

I want to suggest that the LF strcture signals the SM processing John has to do in order to (i) move 'to' the table (meaning move near enough to be able to touch it, i.e. interact with it within his motor coordinate system), and (ii) in this location, performing a reconfiguration action that brings him into a stable, known configuration with the table. It's a sequence of actions. I'm assuming the following:

- V indicates two things. Firstly, selection of a particular (iterated) motor action: walking. Secondly, establishment of a new locomotion environment: (the room John's in)—and the place where John currently is *within* this environment.[5]
- The presence of PathP as the complement of V indicates the activation of the new locomotion environment.
- Path (the head of PathP) denotes selection of a *trajectory* in the newly-established room environment, which is defined relative to a landmark. It doesn't provide much information about the trajectory itself, but the word *to* does convey the *type* of trajectory. Every trajectory is defined relative to an attended **landmark**; different types of trajectory use this landmark in different ways. In this case, the landmark identifies a goal place in the environment.[6]
- [Spec,PathP] denotes
-
-

Now consider a more elaborate PP:

(7.7) John $[_{VP}$ climbed $[_{PathP}$ to $[_{PlaceP}$ on $[_{DP}$ the table]]]]

7.6 A SM interpretation of the preposition 'on'

PlaceP can be headed by words like *on*, *in*. In this section I will think about these concepts a little more deeply. My starting point is that everything in the 'extended PP system' (i.e. all the XPs on the *spine* of an extended PP) describes SM operations taken *within the cognitive medium that denotes environments*. My idea is that none of these things refer to 'objects' as such. Or at least, they don't refer to them *as objects*.

[5] The verb *go* would pick just the second of these. There are languages (like French) where the 'manner' of going has to be reported as a manner adverbial—which indicates that these are separate things that just happen to be read out together in English. Other languages indicate *go* and *walk* morphologically, I think, giving *walk-go*.

[6] The landmark only gives *partial* information about the trajectory, though. There may be obstacles in the locomotion environment, and in this case we're not being told about those. However, we *could* be told, with the more elaborate PP *$[_{PathP}$ [the sofa [round $[_{PathP}$ [the door [to]]]]]]*. *Round the sofa* describes the trajectory that gets to a place the other side of the sofa; this is the *start* place for the next trajectory that gets from this place to the door.

One point to note straight away: my SM interpretation of LF says that when *on* is at the head of PlaceP, it's got to be understood as denoting some *operation.* So now we know *on* denotes an *operation* within the environment medium.

Consider *The cup is on the table. On* appears to describe a relationship between an object (the cup) and an environment (the table). But we're not allowed to refer to objects *as objects*, as just discussed. So how might the environment medium represent the cup? Bear in mind that its representation has to support a representation of the relation denoted by *on*—i.e. a relationship of 'stable support and contact' (see Baillargeon *et al.*, 1998 and much subsequent work).

My suggestion is that the cup is represented in the environment medium as a *collection of environments*—or more precisely, a collection of *surfaces.* (A surface is a kind of environment in itself). This representation is probably computed in the grasp pathway, as distinct from the reach pathway.[7] In the grasp pathway, an object is represented as a set of opposition spaces (see Iberall and Arbib, 1990). That's in a 'motor environment'.[8] But I think there are also 'gravity environments', and there, an object is probably represented as a collection of planes, with an emphasis on flat planes and planes created by triangular elements.

So: how can we define the relation denoted by *on*, at the head of PlaceP? My suggestion is that the relation is defined for an object O and an environment that's horizontal and flat enough that it counts as a gravity environment—or at least that has *places* within it that are gravity environments in this way. I suggest that object O is *on* a surface S when one of O's component planes is coincident with the plane of the S (defined at the *place* in S where O is).[9]

So what's the *operation* involved in establishing the 'on' configuration between an object and an environment? I think it might have to do with *selection* of a surface from amongst the candidate surfaces. Lots of other XPs describe operations where a single item is selected from a set of candidates. For instance in the cup-grabbing sentence, AgrSP and AgrOP describe processes of selecting objects to attend to; VP describes a process of selecting a motor programme. I propose that the maximal projection of PlaceP represents an object as a set of candidate surfaces, and the head of PlaceP describes the operation of selecting one of these surfaces. It's selected because it's the surface that's coincident with the containing environment.

I think a locative PP spine (e.g. *on the table*) describes a process that starts with attention on an object, and then *simultaneously* identifies (i) the containing environment;

[7] It's probably quite useful to separate these two pathways when trying to think of the motor system as a component of the environment system. (In Part 2, there's a discussion of the body as a 'motor environment': see Section 2.19.6.

[8] I also like the idea that the human body is a collection of surfaces. (See again Part 2 Section 2.19.6.) When it comes to defining an agent's relationship with the environment—both with the ground and with barriers like walls—I think these body-centred surfaces are key.

[9] That last bit about place is important: the containing environment doesn't have to be flat, and it doesn't have to have the same surface orientation at every place. A book could be on a cello case: in which case it has to be at a *place* within the cello case environment where the surface is sufficiently flat to be a gravity environment.

and (ii) the relationship between the object and this environment. Recall from Part 2 Section **??** the idea that when an observer attends to an object he's also *obliged* to attend to the environment that this object is in. Otherwise the object can't be indexed to anything. And, I think, the environment the object is in has to satisfy one other constraint: it has to be related to the *current* environment. This constraint relates to the fact that one of the unique things about cognitive representations of the environment is that they have an *LTM* component: every time I change to a new environment, the new environment has to be *connected* to the *current* environment, so that the world remains a connected place. It's not enough to talk about a 'what' pathway and a 'where' pathway: the 'where' pathway must also chart an environment within an LTM structure of known, contiguous environments.

The above point makes me think that the PP system must describe some *continuous* succession of environments—a set of *neighbouring*, or *contiguous*, environments.

Consider when I notice a cat in the garden, out of my office window. (To continue with the example I used in Part 2.) This corresponds to the existential sentence *There was a cat in the garden.* In this example, my attention is initially drawn to a point in a *retinal* coordinate system, that has no proper semantics at all. (This is consistent with the expletive subject *there* denoting 'an as-yet uncategorised point in space'.) My idea is that this location is now categorised in *two* cognitive systems. One is the well-known object classification system, that works out that it's a cat. (A single cat.) The other is the environment classification system. This does something quite a bit more complicated. Firstly, it has to identify two things: (i) an environment, which is represented as a set of places and thus has some spatial structure; and (ii) a place *within* this structure. I'm not sure which of these happens first, or if they happen simultaneously. (My guess is that they happen simultaneously.) In addition to this, the environment selected in (i) is constrained (by the continuity requirement discussed above) to connect to the environment *I'm currently in.* It could be an environment *inside* my current environment: for instance if I recognise that the cat is 'in the corner' or 'on a table next to the door'. But it could also be in an adjacent environment. As in the current example, where the garden environment that the cat is in is *outside*, and adjacent to, the room I'm in. But I think I *always* find an environment—even if it's a sub-environment of my current environment.

I like the idea that in any context, there's a set of **candidate environments** competing to be selected as the *next* environment. When my attention is drawn to a cat sitting on a table in the room, I like the idea that the table-as-environment is competing to be selected, along with all the other objects in the room. Do I have to identify it as a table before I can establish it as an environment? I really think *not.* Epstein's parahippocampal place area (PPA) is an area that categorises/represents places, not objects: and it does so directly. It's known that the area encodes only what's attended to (see e.g. O'Craven *et al.*, 1999). So if my visual attention is directed at a region of the world, we can expect the PPA to identify the environment that's projected to that region: literally, a set of surfaces, that collectively define a map of places. My question is: say my initial environment was 'the room': a floor, with a set of walls as boundaries, defining a set of places, one of which is occupied by the table. If we perceive the table-as-an-environment directly, how is this table environment *related* to the room environment?

Well, part of the answer is in the question: the table is at a location in the room environment. But I also expect that I'm not allowed to establish the table as an environment unless it stands in some determinate, understood, coherent relationship with the room (i.e. with the floor and possibly boundaries). Axiomatically, if the table is a *static* stimulus in relation to the room, there's some stable relationship of support between them.

The PP I'm ultimately trying to understand is *on the table.* But in the above discussion, I'm thinking about *another* spatial relationship: the table being *in the corner of the room.* (And that doesn't even mention the obvious fact that the table is *on the floor* of the room.) In the example I'm considering, the table is introduced with a definite DP, so we can assume that when it was first established, its relationship with the floor/walls was computed. So now, it's somehow available as a *candidate environment.*

The idea of a set of candidate environments is a very useful one. When I attend to an object, one idea is that I always *set up* this object as an environment, even if I don't attentionally enter this environment. For instance, when I first look at the table in the corner, I don't just establish it as the object 'table' at a particular place in the room environment. The PP medium is doing something at the same time, and setting it up as a candidate environment: I think this means (i) that it's represented as a set of surfaces, and (ii) that one of its surfaces is identified as coinciding with the local surface of the containing environment.

7.7 PPs inside DPs: *A corner of the room*

Consider *John went into a corner of the room.* The PP here is interesting, because it has some nested structure. John can easily *start* his locomotion movement already *in* the room. So what's the order of establishment of the representations 'room' and 'corner'? My idea is that *a corner* denotes a *part* of the room, in the way defined above in Section 5.16. You definitely need to establish the whole room first. But having done so, when you pick a place in the room, this place is defined primarily in one *sub-environment* within the room—in this case, for instance, one pair of adjacent boundaries.

Question: why does the DP *the room* appear *late* in the sentence? My assumption is that it appears early first, as a *goal* location, and that the late appearance signals its representation as the *current* location.

So what does the preposition *of* denote? In this case, it clearly denotes some kind of part-whole relationship between an environment and one of its sub-environments. My guess is that it signals an attentional operation that takes an environment, defined as a set of pairs of boundaries, and selects a sub-environment, defined as a *subset* of these pairs. The sub-environment is a way of representing a place within the environment, and equally, a way of representing various trajectories relative to that place.

The question is: what does this operation have in common with other operations that can be denoted by the preposition *of*? This preposition also appears in pseudopartitives like *a can of worms* and 'a line of soldiers'. I think this operation is one in which we establish an object, then re-categorise it as an environment, and then identify something about the *contents* of that environment. (A special kind of contents, in that they in some sense 'fully occupy' the environment—so there's no associated action of *focal* attention.) This is a difference: in *a corner of the room*, we establish an environment and then establish, or attend to, a sub-environment, i.e. a *place*, rather than the contents of this place. (*Corner*, *middle*, *edge* etc are special words that can appear in these positions; they're a closed-class set. We can't say *John went to the dog of the room.* This gives me some hope that these words identify a particular stage of SM processing in the environment/place system.)

Now consider possessive 'of': for instance, *The house of a friend of mine.* Here the object denoted by *a friend of mine* is clearly established as an environment, containing things that are possessed by him; and one of these things is a house. Here it looks like what's happening is some kind of back-pointing: the house is attended to and categorised first, and when this happens, our attention is drawn back to the place this object sits in, which in this case is an abstract space associated with my friend 'as an environment', containing his/her possessions...

7.7.0.0.1 Aside: mass nouns and surfaces One useful idea is that **mass nouns** refer to the material from which *surfaces* are formed.

7.8 A SM interpretation of adjunct PPs

As already discussed, LTM-environment-establishing operations can either happen in the initial context, or they can happen at a specific point in the middle of an episode-apprehending SM routine. In the latter case, they are described by a PP that is the argument of a verb; in the former case, they are described by a PP that's an ajunct to the sentence. For instance *In England, they love real ale.* The latter type of environment-establishing operation is interesting, because it emphasises that PPs make changes to LTM structures: they change the active component of the giant LTM structure that represents the observer's spatial representation of places in the world. The observer can establish an arbitrary part of this structure, in an operation reported in a sentential adverb PP, and all subsequent SM operations will be deictically referred to the newly-established LTM environment. It doesn't have the effect of *placing the observer* in the new environment; that's only allowed if the environment-estblishing operation happens in the middle of an episode-apprehending routine. And in *this* special context, the only environments that are represented as potentially establishable are those that are physically accessible for the observed agent (whether it's the observer himself or an external agent).

7.9 Summary: a SM interpretation of PPs

Chapter 8

A model of episodic LTM

8.1 Background: a model of WM for episodes

8.2 Desiderata for the model of episodic LTM

Here are some desiderata for the model.

One is that episodic LTM must be able to represent episodes as they occur in sequence. Episodic LTM is known to be sequentially structured: subjects find it easier to recall episodes in the order they occurred, rather than in reverse order.

A second desideratum is that episodic LTM should be able to represent overlapping episodes. Not all episodes occur in sequence; we can also remember that one episode was ongoing when another happened.

A third desideratum is that episodic LTM should be able to represent *hierarchical structure* within episodes. Some episodes contain other episodes within them. A classic example is making a cup of tea: this is an episode in its own right, which might follow one episode and precede another; but making a cup of tea requires a number of component actions. The component episodes can be structured sequentially, but there can also be episodes that can occur simultaneously, so the general requirement that episodic LTM can represent both sequentially structured and simultaneous episodes applies to episodes that form components of a larger episode, as well as those that do not.

A fourth desideratum is that episodic LTM should be able to represent absolute times, and identify particular times at which episodes occur. An episode can be located in relation to other episodes (for instance as occuring before, after or during another episode), but in some cases it can also be related to some absolute period of time, such as yesterday, or last December.

A fifth desideratum is that episodic LTM should have a role in planning the future, as well as in representing the past. Evidence has emerged within the last few years that the hippocampus, which is a seat of episodic memory, is involved in imagining the future, as well as in retrieving information about the past (see e.g. Hassabis *et al.*, 2007; Martin *et al.*, 2011).

A sixth desideratum is that episodic LTM should be able to represent the *states* that obtain at different times, both in the past and future, as well as the episodes that occured. This representation of states should interface well with representations of episodes, so that, for instance, the action of heating water in a kettle is performed in a situation in which the water is *cold*, and results in a situation in which the water is *hot*. This requirement is difficult to meet, as it raises the so-called 'frame problem': it is hard to specify which states change as the consequence of an episode.

8.3 Outline of the model

The core idea in this model is that the episodic LTM system represents situations in time using circuits similar to those used to represent *spatial* environments and the objects they contain, as described in Chapter 2 (see in particular Section 2.7).

As summarised in Section 2.5.7, a spatial environment of a given shape imposes a spatial structure on a map of 'places', defining topological or topographical relationships between places in the environment, and implemented in hippocampal place cells. An LTM representation of the environment is a localist assembly in a parahippocamcampal area, that when active, re-imposes the associated spatial structure on the map of places (see Section 2.19.1), and enables a particular set of associations between places in the map and LTM individuals (see Section 2.7.2), so that activating a place triggers activation of the object occupying that place and vice versa. (Of course, since objects can move, these mappings must be conditioned on times as well as on environments. That is one of the issues to be dealt with in the current chapter.) Crucially, any object occupying a place can also be construed as a spatial environment in its own right, that imposes its own spatial structure on the map of places, at which other objects can be located: this stipulation allows spatial representations with hierarchical structure.

I suggest that a circuit similar to this operates for representing episodes at past (and future) times in LTM. The model I introduce here includes two key proposals. One is that *periods of time* are represented in LTM using a circuit similar to that which represents spatial environments. I suggest that periods of time, like 'an hour', or 'a day', or 'today', are represented in a similar way as spatial environments. Just as spatial environments impose a topological/topographical structure on a map of spatial places, I suggest that units of time impose a structure on a map of *temporal* places. The things that occupy these places can be construed as units of time in their own right, just as objects occupying particular locations can be construed as spatial environments in their own right.

The second key proposal is that complex, high-level episodes are also represented in LTM using a circuit similar to that which represents spatial envrionments. I suggest that a complex episode is also something that can be reconstrued as an environment: in this case, an environment which 'contains' *other episodes*. When an episode is reconstrued as an environment, it imposes a structure on a map-like representation, whose locations can be 'occupied by' finer-grained episodes that are constituents of the wider episode—and it modulates a mapping between these locations and episode representations, linking particular sub-episodes to particular locations. The topological/topographical structure of the map of locations encodes alternative 'trajectories' through the component sub-episodes, representing alternative possible ways in the broader episode can take place.

The idea that times and events can be construed spatially is not at all new: it connects to a long tradition in cognitive science, particularly in cognitive linguistics (see e.g. Lakoff and Johnson, 1980 [and some more]). The traditional motivation for this proposal is that spatial language permeates descriptions of temporal relationships between episodes, and descriptions of units of time. The main novelty in the model is in the network architecture that implements the quasi-spatial representations that encode times and episodes, and in links between this architecture and recent findings about hippocampal representations of space and time. I introduce representations of times in Section 8.4, and representations of episodes in Section 8.5.

8.4 A circuit for representing units of time

8.4.1 Timing mechanisms and representations in the brain

The brain has its own internal timing mechanisms, that identify units of time varying in length from seconds to minutes. These appear to be implemented in oscillatory patterns of different frequencies in prefrontal and parietal cortices, that are integrated in the striatum (see Mattell and Meck, 2000 for a review). The resulting patterns can identify a range of time intervals, encoded in a logarithmic scale, with finest accuracy for the smallest intervals, and progressively coarser accuracy for larger intervals, mirroring agents' accuracy in timing tasks (see e.g. Kim *et al.*, 2013). By themselves, these mechanisms do not relate to long-term memory: they are better regarded as part of the medium of perception, or perhaps working memory. However, the representations they create can be referenced by the mechanism that records episodes in long-term memory, so that episodes are associated with particular times. This appears to happen in pathways that connect parietal/prefrontal representations to hippocampal representations. These pathways pass through parahippocampal areas—in particular the parahippocampal place area. (Parietal cortex and the parahippocampal place area play important roles in the representation of space, as discussed in Sections 2.2 and 2.4.1: the fact that these same areas appear involved in the representation of times provides some initial indication that times and places are represented using similar circuitry.)

The hippocampus typically represents units of time in conjunction with representations of space, objects and events, rather than by themselves. However, time representations are certainly a distinct component of these combined representations, as several recent studies have shown (see Eichenbaum, 2014 for a review). An idealised 'time cell' in the hippocampus encodes a particular time, relative to some task-specific start time, in the same way that a place cell encodes a particular location. A population of time cells encode different times in relation to this start time, using the same kind of logarithmic scheme as is found in the brain's primary timing circuits, with times more distant from the start time represented more coarsely (see e.g. MacDonald *et al.*, 2011).

8.4.2 Components of a model of times

A model of the human capacity to represent times must obviously permit the analysis of times longer than those that can be accurately measured using primary timing mechanisms: for instance, it must support representations of time units such as weeks, months and years: episodes are often situated within large units of this kind. In the model I will present, there is an operation that changes the way the way the brain's primary timing mechnisms are interpreted, so that they can encode larger or smaller periods of time.

A model of the human time representation capacity must also allow a representation of evenly-spaced time intervals within any given interval: for instance, a year contains 12 months, and a day contains 24 hours. The model I present has a mechanism for doing this, which co-opts ideas about how numbers are associated with quantities.

Finally, a model of times must represent the distinction between token time units and and types of time unit: for instance, there must be a distinction between a represesentation of a particular week (such as last week, or next week) and a representation of a generic week (which supports expectations about what might happen at particular times within any arbitrary week). The model I present will make this distinction as part of the mechanism that represents time scales.

8.4.3 Architecture of the time network

The structure of the time network is shown in Figure 8.1. I will introduce the media in the network one by one.

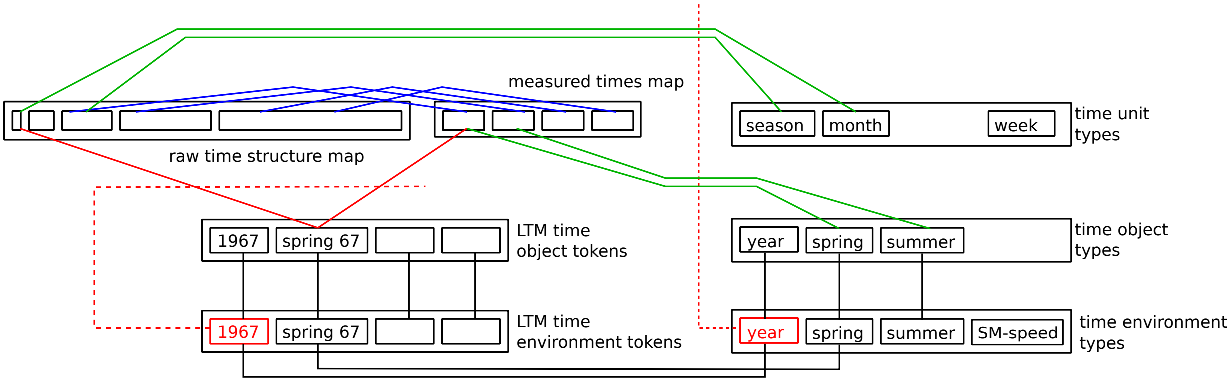


Figure 8.1: Architecture for the network representing token time periods, and types of time period, in LTM

The **raw time structure map** at the top left of the figure holds representations generated by the brain's primary timing mechanisms. I assume these representations are localist, so different assemblies of units represent different units of time. Time is represented logarithmically, as described above: the shortest times that can be measured (shown at the left of the map) are represented with high precision compared to the longest times (shown on the right of the map).

Units in the raw time structure map cannot be interpreted directly: they are interpreted deictically, with reference to the currently active **time environment type**. In experience mode, when the agent is acting or observing, the active time environment type is 'SM-speed': in this scenario, time units have their actual interpretation. But when the agent is rehearsing sequences of episodes retrieved from LTM, other time environment types can be selected. In the scenario illustrated in Figure 8.1, the active time environment type represents a whole year. This means that units in the raw time structure map denote much larger time periods. In the illustrated scenario, the smallest unit of time represents a 'month', and the next-smallest unit of time represents a 'season': these relationships are encoded by associative connections, shown by the green lines running along the top of the figure, linking units in the raw time structure map to units in another medium, that holds **time unit types**. The mapping between raw time units and time unit types is modulated by the active time environment type: this modulatory relationship is shown by the red dashed line crossing the green lines.[1]

The role of time unit types is to permit the measurement of *uniform time intervals* within a time environment. I assume this is a fairly high-level way of gauging time intervals, rather specific to humans. In the current model, it is implemented by a process akin to counting, in which the agent repeatedly waits for a small unit of time, updating a more symbolic representation after each unit elapses. I assume that when a raw time unit fires, it can be *reset*, so that it fires again, without disturbing the mechanisms that cause firing of units encoding longer, coarser-grained time units. I envisage another time map that holds a representation that is updated each time the selected time unit fires: this is called the **measured time map**. When a new time environment is activated, and time units begin to measure time, the measured time map is initialised to some standard value. If time is being measured by a counting process, the selected time unit type will regularly activate and be reset: after each activation, the representation in the measured times map is updated. (I envisage a recurrent network generating the updated representation, based on the current representation and also the active time unit type.) In themselves, the representations in the measured times map do not represent times; they represent points in a discrete sequence. However, associations can be learned between these representations and units in the raw time structure map. These are shown in blue in Figure 8.1. These associations are similar to the associations between numbers and pre-numerical quantity/numerosity representations that children have to learn when first encountering number concepts: in this sense, my model is like the model of number learning outlined by theorists like Lipton and Spelke (2005).

Units in a 'map' of time periods are like places in a spatial map, in the sense that

[1] I think I like the idea that these environment-modulated mappings are stored in a SOM, which receives inputs from the active time environment type, the active time unit type and the active raw time unit. This way, when 'year' is the active time environment, activating 'season' or 'month' will allow activation of the appropriate time-unit type, through top-down reconstruction. But maybe you have other ideas, Martin!

they can be *occupied*. In the current model, the things that occupy periods of time are **time objects**. Time objects are distinct from time environments, but they stand in a 1:1 relationship, in the same way that spatial environments have a 1:1 relation with objects in the circuit encoding spatial representations. Each spatial environment can be reconstrued as an object within a larger environment; similarly, each time object can be reconstrued as a time environment, with other time objects inside it. If the currently active time environment type is 'year', as illustrated in Figure 8.1, then specific units in the measured times map are associated with particular **time object types**. This association is shown by the lower pair of green lines, that connect the first measured time unit to a time object type unit representing 'Spring', and the second measured time unit to a time object type representing 'Summer'. (Note that these associations are also modulated by the currently active time environment type, as shown by the intersecting red dashed line.) Note that a time object type like 'Spring' or 'Summer' can be re-attended to as an environment, by following the 1:1 connections between time object types and time environment types. If this were to happen, a new time environment type would become active. I will talk about this operation below.

So far I have only discussed representations of *types* of time interval. The system of time object/time environment types serves mainly to impose a meaning on units in the time structure maps (that is, in the time structure and measured times maps). However, it is also important to represent *token* periods of time. In the current model, these are represented in the **LTM time object tokens** and **LTM time environment tokens** media, at the bottom-left of Figure 8.1. These media also hold localist representations: assemblies that represent *specific times* in the agent's past. (Or in the agent's future, or even in the agent's present, as I will argue below.) An LTM time environment token links to a single time environment type, as shown by the black lines at the bottom of the figure. When a particular LTM time environment is activated, a time environment type therefore becomes active, imposing a meaning on units in the time structure maps. This token time environment can contain other token times—necessarily shorter ones, of course. This containment relationship is represented by the solid red lines linking units in the time structure maps to LTM time object tokens. Again, these lines represent associations that are modulated by activity elsewhere: in this case, by the currently active LTM time environment token. Thus, when the unit representing the token LTM time environment '1967' is active, there is a mapping from the measured time unit associated with the time object type 'Spring' to the token LTM time object 'Spring 1967'.[2] This time object can also be re-attended to as an environment, in which case the active time environment will change accordingly.[3]

[2] There is also a mapping from 'Spring 1967' to a unit in the raw time structure map, which might represent some less quantitative way of referring to this time, such as 'early in 1967'.

[3] Perhaps the new active time environment would be include 'season', as well as 'Spring'. In this case, there can be multiple time environment types active simultaneously. Note that all time environment types can also be time unit types. But this redundancy is necessary: it's important to be able to activate a time environment type and a time unit type *simultaneously*, so that you can count units within that environment.

As regards how language interfaces with the media in this network, I propose the following principles:

- LTM time object tokens can be read out as proper names: they're a lot like objects. For instance, *D-Day* is the name of a particular token day; *1967* is the name of a particular token year.
- LTM time object types can be read out as common nouns: for instance *last Spring* is an (indexical) reference to a particular token entity of type 'Spring'.
- I suggest that LTM environments, and LTM environment types aren't directly associated with words; to refer to environments, you have to use a PP, for instance *in 1967* or *of 1967.* A PP, like *in X* and *of X*, is an instruction to activate the environment associated with the object *X*—in this case, a time object.
- It's also possible to read out object tokens by reference to their containing environment. For instance, *The Spring of 1967* is read out by identifying a token object of type 'Spring', and then identifying the environment this object is in. (This process is exactly analogous to the process of identifying an object by the object that owns it: for instance, *The sister of Mary*, or equivalently *Mary's sister.*[4])

8.4.4 Representing 'now', and recent episodes

Episodes that have happened recently are stored and recalled better in LTM. There is good evidence that the hippocampal region has a particular role in storing recent episodes, as well as simply creating new long-term memories: damaging this region impairs the ability to create new long-term memories, but also results in retrograde amnesia, that is, the loss of memories of episodes that happened immediately prior to the damage. 'Immediately prior' means different things in different species and different individuals, but in humans the hippocampus appears to provide a temporary store for episodes for around a couple of months at the most.

When an episode is experienced, it has to be associated with a token time unit. Token time units for recently-experienced episodes have to be more readily available, somehow: part of a special medium that represents recent episodes with increased resources. But representations have to *fade* in this medium. At the same time as they fade, the temporal order of representations must be preserved: the effect of these two constraints is that the episodes associated with remoter times are progressively less clearly represented, and after a couple of months have effectively disappeared.

[4] In fact, I like the idea that number expressions, like *twenty-five*, are exactly like this as well. On this hypothesis, *twenty-five* is 'the five in the environment of twenty', and *One hundred and twenty-five* is 'the five in the environment of twenty, which is itself in the environment of one, which is in the environment of hundred'.

8.4.5 States and time units

States are associations between LTM individuals and properties. The properties can be 'simple' ones: location, shape,[5] and things in the RPC, or they can be 'activities', that is, facts about an object's participation in an ongoing episode. An important idea is that these associations are contingent on times, because states can change from one time to another. The relevant times are times at which episodes occur, because changes of state constitute episodes. This means that change-of-state episodes can be defined over many different timescales. (For instance, *Ali grew up* took place over a couple of decades. During those decades, various of my properties were changing, very gradually.)

When I identify a state, I presume I have to associate it with a *time period*. If this extends beyond the present, then this involves making inferences about how long it's been going on, and how long it's likely to persist. Sometimes these are rather precise: for instance if I put the kettle on, I can add that it's boiling, and will continue to do so for another minute or so: I'll have an expectation about the time it will stop being true. These inferences seem very hard; I assume they are the product of lots of learning. The main question is: what are the relevant kinds of learning?

One possibility is that our attention to episodes supplies us with specific training data about how long certain states tend to last. For instance, when we are watching an apple fall, we can learn how long an apple takes to fall; similarly we can learn how long a kettle takes to boil, how long a person needs to tie his shoelaces, and son on. We can also learn about the time that states obtain: for instance, how long an orange stays fresh, how long an orange stays in the fruit bowl, how long a child stays young. Then when we identify a state, we can store it in (semantic) LTM associated with a time unit of a suitable duration. In cases where the state is relevant, we can associate an action checking the state with a future time, so that when that future time is 'now', it's an action we consider.

Another possibility is that I only represent certain states when I'm experiencing, or simulating, a structured sequence of events. That is, I may only represent certain states within a framework where I experience, or relive certain structured experiences. I like that idea. But it can't always be like that: I can say *When I was a boy* to retrieve a context from LTM, without any special framework of episode sequences, so the association between me and the type 'boy' must serve to retrieve a specific LTM time period. (Maybe what it retrieves is a set of LTM times. In this case they'd all have to be times at the same level of temporal granularity, of course.)

8.5 A circuit for representing LTM episodes

Units of time can contain episodes. Episodes transform the properties of objects, and can also create and destroy objects. However, it is problematic to associate episodes directly with absolute units of time. This creates conundrums: at exactly what time do I start or

[5] An object's shape is identified by establishing it as an environment, and looking at the spatial structure of the map of places (the 'allocentric boundary structure'), as discussed in Section 2.11.

finish grasping a cup, or making a cup of tea? Semanticists often avoid such conundrums by representing time *in terms of* episodes. In the situation calculus, for instance, we refer not to absolute times, but to the temporal context brought about by the completion of particular episodes (see e.g. Levesque *et al.*, 1998 for an introduction). I suggest that there are two systems for locating episodes in time: episodes can be situated roughly at absolute times, but can also be situated more precisely, and more 'semantically', in relation to each other: for instance, an episode can be represented as occurring after some other episode, or as occurring within some other episode. Studies of hippocampal representations certainly provide some *prima facie* support for this idea: the firing of hippocampal cells is often modulated by time, as discussed above, but is also often modulated by the animal's task (see again Eichenbaum, 2014).

In this section I will introduce a circuit for representing episodes in relation to one another. I'll call it the **LTM episodes network**.

The architecture of the episodes LTM network is shown in Figure 8.2. In this section

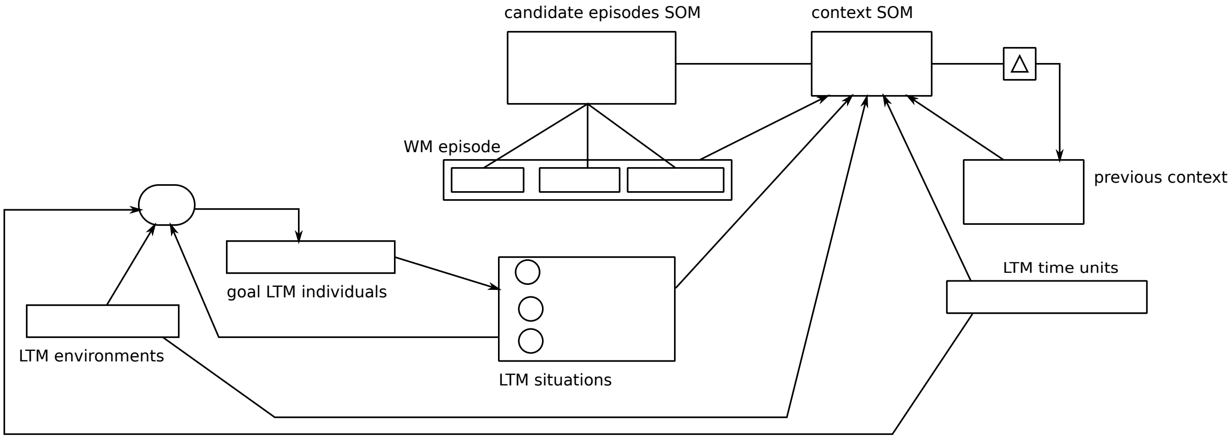


Figure 8.2: Architecture for the network representing episodes and situations in LTM

I will briefly introduce various different aspects of the architecture: each will be discussed in more detail later in the chapter.

8.5.1 Expanded roles for the candidate episodes SOM and current context SOM

In the model of working memory introduced by Takac and Knott (the Cognition paper), the candidate episodes SOM held a probability distribution over expected or rewarding episodes in a given context, that provided a real-time top-down bias on the agent's experience of the current episode. This distribution was generated by the current context SOM, as a function of the sequence of episodes that had been experienced. In the current model, these SOMs still have this function of generating top-down biases on incoming experiences. However, they also have a wider role in a model of episodic LTM.

I now propose that the candidate episodes SOM holds not only a distribution of expected or useful episodes, but representations of *all episodes that are stored in the agent's episodic LTM*. And similarly, that the context SOM holds not only all *types* of context that the agent can encounter, but also all *token* contexts that he can retrieve memories about. On this model, retrieving an episode from episodic LTM involves activating a particular context in the context SOM, by some means, and then generating a distribution in the candidate episodes SOM as usual: but now this is a distribution over *what actually happened* in the specified context, and the winning episode is an episode 'retrieved from LTM'. I will discuss how episodes are stored in and retrieved from LTM in Section 8.12.

The candidate episodes SOM and context SOM should now be thought of as very large media: the candidate episodes SOM holds large numbers of localist episode representations, and the context SOM holds large numbers of distinct context representations. Of course, since these media are both SOMs, they are able to generalise if need be. So, for instance, the candidate episodes SOM does not have to store every single episode that has ever been experienced as a token: it can also hold episodes at varying levels of genericity. Some episode representations will involve token LTM individuals, while others might just refer to individuals with particular properties. Likewise, the contexts SOM can hold representations of some token contexts, if they are very salient or important, or commonly revisited. But it can also hold representations of types of context. It's useful to have SOMs holding these representations, because they can learn generalisations if they need to. Nonetheless, both these SOMs are significantly larger than they would be if their role was just to store expectations about forthcoming episodes.

The idea that the same circuits that compute representations of expected episodes are also involved in holding representations of remembered episodes makes sense for several reasons. For one thing, it fits well with a constructivist model of LTM recall: when we are remembering what happened in a given situation, it is well known that we make use of generic knowledge about what *normally* happens in such a situation (see e.g. ??). For another thing, there's interesting recent evidence that imagining or planning episodes in the future involves the same mechanisms as are involved in representing episodes in the past (see again Hassabis *et al.*, 2007; Martin *et al.*, 2011.

There are also some useful technical advantages to this architecture. For example, if remembered episodes are picked from a distribution of possible episodes, there is an associated measure of confidence: we might be very sure about the episode that occurred, or very unsure. This measure of confidence could feature in an account of modal statements about remembered episodes (for instance, 'E *may* have occurred', '*Perhaps* E occurred'). As another example, the architecture also allows a nice treatment of counterfactual reasoning. When retrieving from LTM, we generate a distribution over episodes, of which the most active item can be taken to be the episode that actually occurred. The other items can be taken to be episodes that *could* have occurred. In the current model it is quite straightforward to simulate the occurrence of one of these episodes, and generate a representation of the context that would have obtained, and of the distribution of likely episodes in that context, and so on.

It should be borne in mind that even in the working memory model introduced in

Takac and Knott, the candidate episodes SOM and the current context SOM both hold representations learned over the course of long experience, and are therefore already part of the 'long-term memory' system. Their only contribution to working memory is in the assemblies that are currently *active* within them. (This makes them different from the WM episode medium, which is a genuine working memory medium, where SM representations are held active for short periods of time, and where there is no long-term learning).

8.5.2 A multi-faceted representation of context

Given that the context SOM now has a role in representing remembered contexts, as well as just types of context, it must now have a much wider range of inputs. A key idea in the current models is that there are several different cognitive representations of an agent's 'current circumstances', and that these are combined in the context SOM. This SOM is updated after each episode is experienced: it takes as input the current WM episode, as well as a copy of itself at the moment before this episode was experienced (the **previous context**): that is, it is a recurrent SOM (citation), that is able to store sequences of episodes. However, it also receives a number of other inputs, allowing these sequences to be conditioned on several other factors. Two of these inputs have already been introduced: the context SOM takes input from a representation of the current **time** (see Section 8.4) and a representation of the currently active **spatial environment** (see Section **??**). The other input is new: the context SOM also takes input from a representation of the 'current situation'. In the next section, I will briefly introduce this representation.

8.5.3 LTM situations

The **current LTM situation**[6] is a sparse, localist assembly, representing the agent's current cognitive set. That is, it's a representation that doesn't update very easily, being relatively immune to distraction: the kind of representation that's held in prefrontal cortex in many models (see e.g. Miller and Cohen, miller2001; Braver and Cohen, 2000).

The LTM situation is most easily understood as representing a currently ongoing task that the agent is performing. But it can also be used to represent routines, or scenarios. An agent's 'current circumstances' also reflect the episode that has just been experienced, and, more generally, the circumstances that have recently obtained.

Importantly, the LTM situation presents a relatively tonic input to the context SOM. While this SOM's activity is updated after each experienced episode, the input from the current LTM situation normally remains unchanged. There are, however, ways of transitioning to a new LTM situation. Such transitions happen when the agent embarks on a new task. I will discuss situations bit by bit: I introduce one type of situation, associated with a plan for making something, in Section 8.7, and I discuss a variety of different types of situation in Section 8.10. [An important thing to say somewhere about situations is that

[6] I later refer to LTM situations as 'tasks', and to contexts as situations. I prefer this terminology: I need to standardise, sorry!

they're by their nature generic. However, there's also a notion of a token situation, which is an instance of a (generic) situation occurring at a particular token time.]

Situations are in some ways like the LTM spatial environment representations introduced in Section 2.19.1. As discussed in that section, the currently active LTM environment imposes a structure on the map of places, and thereby defines a set of potential trajectories (i.e. sequences of places). In a similar way, the currently active LTM situation imposes a structure on sequences of *episodes*, by providing tonic input to the current context SOM. The current context SOM delivers a set of possible episodes to the agent, which is updated after each episode is experienced. This is formally very similar to the representation of the agent's spatial environment, which at each point delivers a set of possible places the agent can move to—a set that is updated after each action the agent makes. Crucially, in the spatial model, there are actions that update not only the agent's current location, but also the agent's whole environment, thereby causing a significant restructuring of the map of places, as discussed in Section 2.18. In the current LTM model, these are analogous to operations that transition relatively discretely from one situation to another, and dramatically change the set of supported episode sequences.

I want to take the analogy between LTM situations and spatial environments quite seriously. I want to argue that it's implemented in the same kind of neural circuit. I also want to argue that it's no coincidence that we use spatial language to talk about situations: people find themselves 'in' situations, or 'in the middle of' situations, for instance. I'll say more about this in Section **??**.

8.5.4 Representations of goals, and hierarchical structures of episodes

Another key set of representations in the architecture relate to an agent's goals. The architecture described thus far allows top-down biases on episodes and individuals, but it does not allow the kind of *persistent* top-down biases that are characteristic of an agent attempting to find a particular kind of object, or bring about a particular state. In the current model, a key idea is that goals are represented in media parallel to those that hold WM and LTM representations of individuals and their states. A key representation is the **goal LTM individual**: a LTM unit similar to the LTM individual introduced in Section 2.7.2.1, except that it holds a representation of a *desired* individual, with *desired* properties rather than an actual individual with actual properties. Goal individuals and their properties will be discussed in Section 8.6, and a more detailed circuit diagram will be provided. The important thing to note in Figure 8.2 is the function that maps the current LTM situation, the current LTM environment, and the current LTM time, onto a set of candidate goal individuals. Note this function does not take input from the current context SOM, which updates after each perceived episode: only from representations that update more slowly. Like the activation of situations, the setting of goals is something that is relatively immune to distraction from ongoing episodes.

In Figure 8.2 there is also a link from goal individuals back to LTM situations. The activation of a goal individual, or goal state, is one of the circumstances that allows the activation of a new LTM situation, as I will discuss in Section 8.10. Since situations implicitly define complex sequences of episodes, this mechanism is one that allows a measure of hierarchy within episode representations.

8.5.5 Situations and time units

'While the kettle is boiling, put some tea in the teapot'. How can I represent this relation of temporal simultaneity? In particular, how do I represent the ongoing episode 'the kettle is boiling'?

In my current model of the progressive, the observer attends to the kettle, and then has his attention drawn to an interesting stative property of the kettle, namely its participation (as agent) in a boiling episode. That works fine if the observer is newly observing this ongoing episode. But another way of arriving at the same awareness that 'the kettle is boiling' is by having observed that the kettle-boiling episode has been initiated, but is still under way. Ideally, this process would lead to the same representation as the property-observing process. The property-observing process activates this representation by observing the kettle in the middle of boiling. The initiation-based process activates this representation as a side-effect of recognising the initiation event.

After observing an activity and registering it as a property, the observer adds a state representation, which is associated with a time period, as discussed in Section 8.4.5. I think something similar is true for the initiation-based process. After having observed the episode that *starts* the kettle on its journey towards boiling, *time has to pass* before it reaches its consequent state of being at 100 degrees. It's during this time that 'the kettle is boiling'. A key point is that this consequent state is relevant for the current task: there's something that should happen afterwards. The association with a time unit in this context can serve a specific SM *function*: there's a natural time when we look for, or expect, the consequent state of the episode to be reached. If we register that the kettle has just started boiling, I suggest this is like *setting an alarm for a future time*, when we expect the kettle to boil. Note that the time representations described in Section 8.4 do involve internal timing devices. (Note also that this timing-device initiation can also apply when the observer happens to notice the kettle boiling.)

The idea of an expected state also crops up in the model of situations. When you're making tea, you're waiting for the state in which a cup of tea is present, and this state triggers de-activation of the situation.

8.5.5.1 Starting and stopping episodes

In this section I'll discuss meta-level episodes that refer to episodes starting or stopping: for instance *The kettle started boiling / started to boil*. I'll argue that these can be assimilated within my general proposal, whereby complement clauses (including these nonfinite ones) describe a process involving a sequence of two consecutive representations in the WM episode buffer: firstly an episode in which the kettle starts, and secondly an episode in which it boils. Crucially, I envisage a mode-changing operation after the first episode.

In detail: I suggest that what the observer first notices is that the kettle begins *doing something*, and this happens before the nature of this something is identified. I suggest this observation puts the observer into a special mode where the thing being done is to be identified as an activity, rather than as a completed episode.

Compare *The kettle started boiling* and *The kettle boiled*. The latter sentence is a complete event: there's no point at which the observer records the activity of the kettle as a stative fact. The main difference is that when the observer apprehends *The kettle boiled*, his attention *remains* on the episode until it's complete: his attention is *held* by the episode, and his update from the WM medium to the LTM medium happens at the point when it terminates. Whereas when he apprehends *The kettle started boiling*, he records something about the episode in LTM prior to the completion of the episode, leaving him free to attend to other things. I also like the idea that at this same time the observer starts a timer running, which can interrupt processing around the point when the episode is expected to finish. This timer-starting action also links the activity associated with the episode to a time in LTM.

How does the observer record *The kettle started boiling* in LTM? The whole thing is an episode in its own right. I propose that the key operation is the mode-setting operation. The observer first attends to the kettle (perhaps because of the activity that has just begun—but that is not yet represented as an activity, just as a cue to salience, because we have to attend to an object first). Then there's a special *perceptual mechanism* that identifies that an activity has just begun. Say it's triggered by the *onset* of an action representation. (This onset event could be the same stimulus that made the kettle salient, though it doesn't have to be.) I like the idea that this mechanism is in the same medium as first-order action representations. (This is why the word that denotes it, *start*, is a verb.) Of course, if an activity has just begun, an action representation is *also* active. In this example, for instance, we have two actions active: one is the first-order action *boil*, the other is the meta-level action *start*. I like the idea that these two are *in competition with one another*. If the winner is *boil*, the observer will see the boil episode through to its completion. If the winner is *start*, the observer stores this in LTM right away, then immediately moves into a new mode where the next thing that will be identified is an activity, and then records this activity in LTM. Now his WM episode buffer is free to experience other episodes *while* the kettle is boiling. If the winner is initially *boil*, but the episode is not completed soon enough, I suggest the observer backtracks and looks for another action—and that this might well be *start*—at which point, *boil* can get back in.

Why do we have a nonfinite clause the second time around? I suggest it's because the original episode has been inhibited in the WM episode medium, to make way for the new one—but at the point this happens, the observer is *still attending to the kettle*.

8.5.5.2 Volitional start meta-episodes

[This section is superseded by stuff about scenarios in Section 8.11.]

The above account is from the perspective of an observer watching a nonvolitional event. However, start episodes can also describe volitional events: for instance *John started to make the tea*. My original idea about this sentence is that it describes John entering the tea-making situation. How does this square with the model just outlined? What event does 'John started' denote here? I assume the observer first attends to John. (Say he is John.) I'd normally assume that the next operation is to activate the goal individual 'the tea', and then the special action 'make' that operates on this goal individual to activate the associated tea-making situation. But in this particular case, John *interrupts* his tea-making. (That's why we don't get the telic sentence *John made the tea*.) John has created the tea-making situation; he may also have associated the activity 'making tea' with a time period, at the end of which there's a reminder to check a goal state.

I suggest what happens is that John activates the WM episode 'make tea' and gets going, but doesn't store anything in LTM yet, because he'd normally only do that when he was finished. But now something happens to interrupt him. At this point he needs to save his current activity in LTM, as an activity (since it's unfinished). Technically this can be done by temporarily activating a very strong representation of the meta-action 'start', which takes over the WM episode. The WM episode *John start* then inhibits itself, and sets up a mode where the next-most-active WM episode will be registered as a property of John, that applies at the current moment.[7]

Actually I like the idea that with some simple actions, once they have been started, they can continue reasonably automatically while the agent allocates his attention to other things. In such cases, it may be that the agent re-attends to the ongoing action when it stops. For instance, say an agent starts running on an exercise machine. While he runs, his WM episode buffer can become occupied by other things: thoughts, plans etc. During this time, however, there will be a tonically active unit representing 'run'. I suggest this allows him to periodically re-establish his running episode. (As a stative activity, I presume.) When he decides to stop, I suggest he must first activate 'stop', which triggers activation of another special mode, in which the WM episode established next is constrained to be a stative activity, and also constrained to be turned off. (The perceptual equivalent of this is where the observer activates an action representation tonically—for instance the humming of a fridge—and then has his attention drawn to this representation by a unit that identifies a change in it. Again, the humming activity is transiently re-activated after the change is noticed, so that the time period of the humming can be properly delineated.[8])

[7] This story is a bit similar to the one I have about how an agent decides to execute a 'talk' action instead of trying to do something himself: his first-order goal is temporarily overridden by the 'talk' goal, which puts him into verbal mode and inhibits itself, letting the first-order goal back in, but in a mode where it will be the content of a speech act.

[8] Note that we could also say *The humming stopped*, where the humming stimulus is treated as an individual—in fact as a possession of the fridge. But here, *stopped* means the same as *ceased*: whereas if a physical object stops, it doesn't cease (one of its actions ceases).

8.5.5.3 While John was making tea, ...

[This section is superseded by stuff about scenarios in Section 8.11.]

Let's say John does get interrupted while making tea. I suggest this requires him to represent tea-making as an activity, rather than a completed state. (Otherwise he'd keep going until he was finished, and then represent *I made the tea.*) When interrupted, he has to save his current activity in LTM in a format that allows him to return to it. In my model, this involves associating it with a time. (Since in my model, dual tasks are performed by keeping a bunch of episodes or situations associated with 'now'.) What's associated with 'now' must be a representation of a volitional episode—otherwise it couldn't be resumed. But it can't be a representation of a completed episode. Those aren't the sorts of thing that are associated with time intervals. So what he associates with 'now' is a stative representation of 'what he is currently doing'.

8.6 Goal individuals and goal states

In the default attentional setup described so far, the observer's attention is biased towards individuals that are likely to participate in the episodes that occurred (or were rewarded) in this situation in the past. This bias is expressed in terms of RPCs, and then converted to locations, via the candidate WM individuals buffer. Importantly, once the bias is imposed, the observer selects the most salient location, and categorises the individual at this location, and goes on to experience an episode involving this individual, without scope for backtracking. However, if the observer is the agent, there is a different mode of behaviour which needs to be modelled: one where the agent decides in advance what object he would like to identify, and *persistently* acts with the aim of establishing this object's presence.

The clearest instance of this kind of persistence is in **visual search**, where the observer's explicit goal is to identify an object of a given type in the visual field. Many models of vision focus exclusively on visual search. But of course an observer in the world is not always performing visual search: sometimes he is just experiencing episodes. The process of attending to and classifying an object is sometimes subordinated to a visual search task. But at other times, it is simply a component of episode perception: the observer attends to the most salient object, and then proceeds to experience an episode involving this selected object.

What determines whether visual attention is subordinated to a visual search task or to episode perception? I suggest this is another place where the concept of a cognitive mode can be helpful. I propose that an observer can be in two modes: one whose aim is to experience the most relevant episode occurring now, and another dominated by an activated *attentional goal*, which establishes a special kind of perception mode, geared towards perceiving *states* rather than episodes. This perception mode relates to episode-perception mode, naturally. But it also relates to action-execution mode, in that it often identifies the preconditions for actions performed by the observer. In this section I'll introduce a model of attentional goals, and the perceptual mode that is triggered by attentional goals.

Experimentally, the kind of mode-switching operations I'm thinking of are related to those discussed by Corbetta and Shulman (see e.g. Corbetta and Shulman, 2002). [More here.]

A key assumption of the model is that attentional goals are selected as a result of structures residing in long-term memory. I propose that an observer in a given situation represents not only the real individuals it contains, but also a set of **goal individuals**. These individuals are akin to LTM individuals: each goal individual is a sparse, localist structure, associated with a rich property complex (RPC). But they occupy a parallel medium: while LTM individuals activate patterns of activity in the 'location' and 'RPC' media, goal individuals activate a pattern of activity in parallel **goal location** and **goal RPC** media. The RPC medium holds the properties of a search target: I envisage a match operation that compares the pattern in the 'actual' RPC medium with that in the goal RPC medium. I also envisage that a selected goal RPC can bias visual search towards objects with relevant low-level visual features, and towards particular locations, just as the actual RPC can. (The goal location medium holds the goals of locomotion actions of various kinds, of the kind described in Chapter 5. A goal location only becomes active if a goal individual is successfully matched to an actual individual, as will be discussed in Section **??**.)

In the model, goal individuals are activated as a function of the observer's current circumstances, via associations learned through reinforcement. The 'current circumstances' include the current spatial environment, the current token time and time type, and the currently active 'LTM situation'—a term which I will discuss more below. At any given moment, these current representations induce a distribution over the set of goal individuals, to create a goal-based analogue of the saliency map. At the same time, a regular saliency map is computed, as a function of bottom-up visual inputs, as biased by the candidate WM episodes medium. I propose that these two saliency maps compete, to determine whether the observer enters **episode-perception mode** or an alternative mode that I will term **task-execution mode**. In episode-perception mode, goal individuals play no further role. In task-execution mode, the observer's behaviour is controlled by the collection of active goal individuals, as I will describe below.

Goal individuals are part of the LTM system: I envisage them being very much like remembered token individuals. This extends to the duality between objects and environments described in Section 2.18.2[9]: as well as a goal individual medium, I envisage a **goal environment** medium, allowing goal individuals to be re-established as environments, within which other goal individuals are potentially located (at goal locations).

The circuitry associated with goal individuals is shown in Figure 8.3.

8.6.1 A visual search scenario, featuring a simple goal individual

A simple scenario is where there is a single goal individual. Say, for instance, that the agent walks into a kitchen, and this triggers activation of a goal individual associated with

[9] And hopefully other places!

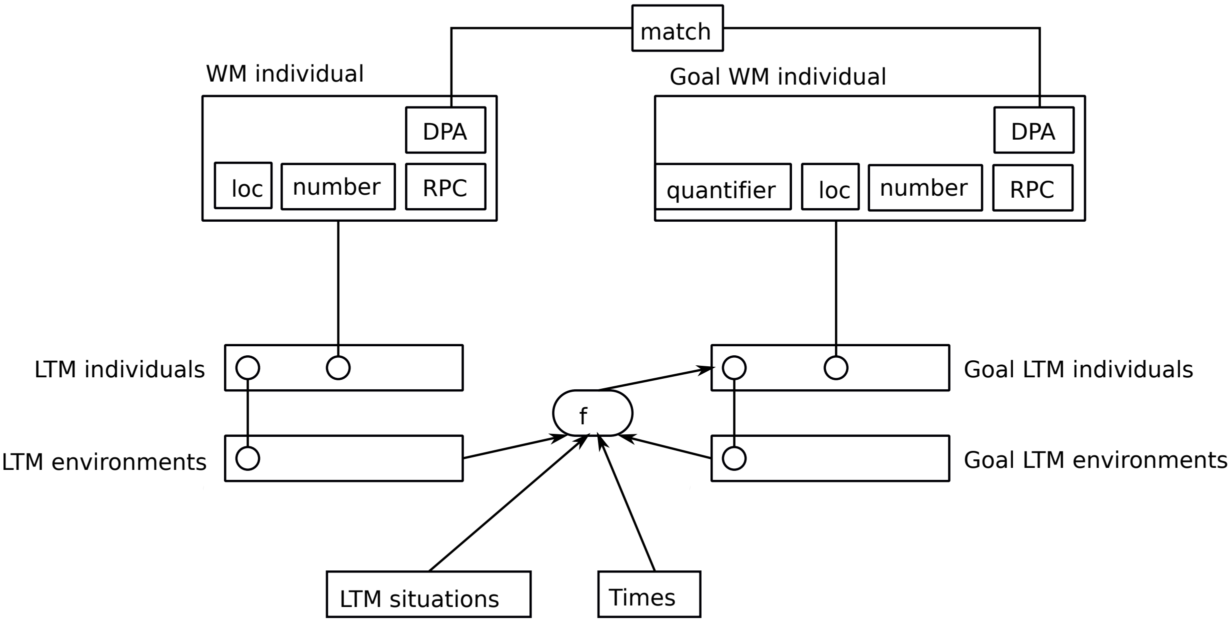


Figure 8.3: Architecture for the network representing goal individuals

the type 'cup', and that this goal individual is sufficiently active to trigger goal mode. In these circumstances, the agent's saliency map will be dominated by objects possessing the low-level visual features of cups, and by locations where cups have recently been (as recorded in the WM individuals medium) and by locations where cups are expected (as recorded in object location LTM). The agent will pick the most salient location, which may or may not correspond to an actual cup.

In either case, the actual RPC is compared to the goal RPC. If there is no match, the currently attended location is inhibited, and the next-most salient location is attended to, resulting in a sustained search for an object matching the goal individual. I envisage that this search can be widened beyond the environment that triggered activation of the goal individual: for instance, if the agent cannot find a cup in the kitchen, he may look beyond the kitchen. Ultimately there must be conditions on a goal being abandoned: these will presumably relate to a computation of the amount of effort expended, and/or the amount of time since the goal was posted. (Perhaps the activation of a goal individual be slowly decreased over time, by an amount relating to the amount of effort expended.)

If there is a match between the goal individual and a perceived object, the goal individual will also be inhibited, because the search goal is achieved. But the purpose of finding an object is frequently to do something with it: so before the goal individual is inhibited, the identified individual is copied to the 'patient' slot of the WM episode, creating the conditions under which the agent can select an action to perform on the individual. The candidate actions that compete for execution include actions that pick up on differences between the properties of the goal individual and those of the matching actual individual. For instance, if the goal individual has a (goal) property that differs from the (actual) properties of the matching individual, this discrepancy defines the preconditions for a causative action that brings about a relevant change of state in the actual individual. This scenario will be discussed in Section 8.6.3 for inherent properties, and in Section 8.6.4 for location properties.

Comparison to other models of visual search There have been many models of how visual search can be biased towards some particular target object. A prominent model is that of Navalpakkam and Itti (2005). In this model, visual search is controlled by two saliency maps. One represents 'bottom-up' salience, implemented by various forms of local contrast (Itti and Koch, 2000). The other represents task-relevance: this is implemented by a function that biases salience computations towards the low-level visual features of a selected target object. This model is explicitly a model of visual search: if the selected target is not identified, search continues. However, in my model, visual search is only one of the functions of visual attention: it also has a separate role in perceiving episodes. (Including episodes in which th eobserver is the agent, of course.) It is for this reason that I envisage a task-execution mode that is distinct from episode-perception mode, and envisage visual search as being handled within task-execution mode.

In fact, Navalpakkam and Itti's operation of combining bottom-up and task-related saliency maps is already implemented in our model, in regular episode-perception mode. In this mode, the saliency map is influenced both by bottom-up salience and by biases from expected (or rewarded) episodes in the candidate episodes SOM. However, in our model, there is no *persistent* activation of a search goal in this mode. We envisage that persistent activation of a search goal, triggering a sequential and possibly sustained search for a particular goal object, occurs in a separate cognitive mode, namely task-execution mode.

8.6.2 Multiple goal individuals

There can be several goal individuals active in any given situation, so when one goal individual is inhibited, whether due to success or to time-out, another one might become active. The effect of this is that when the agent enters a situation, he may check for the presence of several individuals, to make sure they are the way he wants them to be, and to take action if they are not. Of course, this checking process only happens if there are no actual salient objects or episodes that claim the agent's attention.

The idea of multiple goal individuals also connects to Navalpakkam and Itti's (2005) model. Their model includes a 'task-relevance map' of locations that are relevant 'top-down' to the agent's goal, alongside a regular bottom-up saliency map that detects salient stimuli in the image. Again, the saliency map is more directly linked to our saliency map, since the top-down episodes that modulate it might relate to several alternative tasks. However, our set of goal individuals in some way corresponds to the set of task-relevant entities in Navalpakkam and Itti's model.

8.6.3 Goal properties and causative actions

Often it's not only important to identify an individual of a certain type, but to ensure that the identified individual has certain properties. In this model, the agent is not only alert to possible episodes in the current situation, and ready to act in the situation, but is also *checking* various stative aspects of the situation, to see whether everything is the way it should be. In this section I'll consider inherent properties of a goal individual: for instance, the goal that the sink in one's kitchen is clean.

My main suggestion is that goal individuals can have properties specified in their RPC, as well as properties that define a basic-level type. In the model I envisage, when a goal individual is matched with an actual individual, it is *only matched by type*: that is, a matching item in the real world does not have to have the properties of the active goal individual, provided it has the same type. Once a matching item has been found, the agent can consider the goal individual's properties. My idea is that the operation of finding an actual individual whose type matches that of the goal individual is a little like the operation of classifying a perceived individual: it's only after you've done this that you can pay attention to the individual's properties. In the model I'm proposing, the operation of attending to a goal individual's properties is a bit like the perceptual operation of noticing the properties of a perceived individual that has just been classified: you inhibit the type, and see what's left.

Say that the agent when arriving in a kitchen activates a goal that the sink is clean. This is implemented by a goal individual associated with type 'sink' and additionally, the property 'clean'. This goal individual will first prompt the agent to find a sink. If he finds one, he's then in a position to identify what properties this sink is supposed to have. This involves activating a goal property in the 'goal RPC' medium. This may highlight a difference between the actual sink's actual properties and the desired goal property. For instance, the actual sink may not be clean. If so, we have the preconditions for a causative action to perform on the sink: namely the action that will cause the sink to *become clean*.

Recall that when a goal individual has been processed, the matched actual individual is copied to the 'patient' field of the WM episode buffer, setting the stage for the agent to select an action to perform on the matched individual. In this case, the natural action to select is the causative action 'make clean'.

This discussion of goal individuals fills in a piece of the story that was missing in Lee-Hand and Knott's (2015) account of causative actions. Lee-Hand and Knott specified how an agent learns to perform actions that achieve particular effects: for instance, the action that causes a lever to become bent, or a door to become open, so that if the agent is given the *goal* of achieving these effects, it can execute an appropriate action. But it did not specify how such goals arose in the agent. The current account of goal individuals goes some way to answering this question. (Of course, the question of why a particular goal individual would become associated with a given situation has still not been answered. I will discuss this in Section 8.6.7.)

8.6.3.1 Opportunistic activation of a goal individual

Say an agent has a goal for the sink to be clean, but it's not strongly activated to force the adoption of task-execution mode, and the agent enters experience mode instead. But now assume the agent attends to the sink for some other reason related to experience mode: maybe it's salient, or there's a relevant episode involving the sink to monitor. And assume the sink is in fact dirty. I suggest that when any individual is attended to, this increases the activation of any goal individual that matches it, to make the agent predisposed to check goals related to any individual he encounters. In this case, for instance, the agent might interrupt his experience and notice that the sink is dirty, and generate the preconditions for an action to clean it. I will call this mechanism 'opportunistic' activation of a goal individual.

8.6.4 Goal locations and locomotion actions

The 'goal properties' of a goal individual can include intrinsic properties, but also properties to do with the individual's location. For example, imagine that in the kitchen, the agent requires the teatowel to be in a certain location L. This can be implemented by a goal individual associated with the type 'teatowel', and with the location L. This individual will first prompt the agent to find the teatowel: having found this, the agent will activate L as a **goal location** for the currently attended teatowel. The notion of a goal location has already been introduced in some detail in Chapter 5, in Section 5.1.5: as discussed there, the combination of the actual location of this teatowel and the active goal location create the preconditions for a causative action whose effect is for the teatowel to move from its current location to L. (I suggest this is reported in a sentence like *A moved the teatowel to L*, whose LF is *A caused [the teatowel moved to L]*, as discussed in Section 7.)

As a more complex example, say the agent wants the clean teatowel in location $L1$, and the dirty teatowel in $L2$. In the current model he would activate two goal individuals linked to the type 'teatowel': one linked to the property 'dirty' and the location $L1$, the other linked to the property 'clean' and the location $L2$. If there is no more pressing business (for instance, no episode to monitor), he would enter goal mode, and select one of these goal individuals to control his current behaviour. Say he selects the 'clean teatowel' individual, and the resulting visual search leads to him establishing an actual clean cup. Upon establishing this object, he will identify it as a basic-level type, namely 'cup'.

There's a problem here: we can't distinguish whether the goal is for the teatowel object at $L1$ to be clean, or for the clean teatowel to be at $L1$. Those are very different conditions to meet! My guess is that you just can't specify complex goals like these—that is, goal individuals are defined by their type, and nothing else. (It may still be possible to have a type that's complex enough that there's no word for it, and therefore a relative clause is needed to convey the type.)

8.6.5 Hierarchically structured goal individuals

I'll now introduce a slightly more elaborate goal individual, that defines the preconditions for a slightly more complex locomotion action. Consider a component of a tea-making scenario: one sub-goal in this task is for the kettle to contain a certain amount of water. I'll suggest that this goal can be represented by a goal individual with some *hierarchical structure*.

The goal individual in question is a kettle: so the agent's first objective is to find a kettle. Having found the kettle, the agent can now consider the (goal) properties of the (goal) ketttle, as just discussed in Sections 8.6.3 and 8.6.4. However, in this case, the relevant property of the kettle relates to its contents, rather than its intrinsic properties or its location.

As discussed in Section 2.18.2, every LTM individual can potentially be re-established as a spatial environment in its own right—and this is what we do in order to inspect the component parts of an object, or objects contained within (or possessed by) an object. I propose that goal individuals are just the same: that is, there is a medium holding **goal environments**, alongside the medium holding 'actual' environments, and each goal individual can potentially be associated with a single goal environment.

When the agent 'attends to the goal individual', to see if there's anything distinctive about it as an instance of its type, I propose that he not only inhibits the type of the individual, to see if there are any idiosyncratic properties that remain, and considers its goal location: he also considers whether the goal individual is represented as having any unusual contents, or parts. This involves establishing the individual as an environment, and seeing if, in this new context, any *new* goal individuals arise. This is a recursive operation: when the agent enters the kitchen, a set of goal individuals become active, and now, when the agent 'enters' the kettle, another group of goal individuals can potentially become active. In the current case, there is just one active goal individual, associted with the type 'water'. This individual has a location property: it should be in a particular spatial region within the kettle. (Recall that when the agent attentionally enters the kettle, his cognitive map is restructured to represent places within the kettle, rather than places within the kitchen.) The active spatial region identifies how much water should be in the kettle.[10]

At this point, the agent has a new search goal: he needs to find some water. As noted in Section 8.6.1, this can involve looking outside the currently active spatial environment, in the containing spatial environment—in this case, the kitchen. The place where water is located is in fact within another sub-environment in the kitchen: namely *in the tap*. The preconditions now exist for a causative action in which the water is moved from its current actual location (the tap) to the goal location (inside the kettle). In the default case, this action will be to put the kettle under the tap and turn the tap on.

[10] Note that the 'water' individual is a *mass* individual: that is, it will have 'plural' number, but it will still be associated with the type 'water'. This type is the type of its texture elements, rather than associated with the global form of the individual.

8.6.6 Quantified goal individuals

An agent's goal might relate to a single object, or to a group of objects of a given type. There is a simple way of differentiating: a goal individual can be associated with a cardinality: either singular, when the goal relates to a single object, or plural, when the goal relates to a group of objects. But in the latter case there are additional degrees of freedom. We might want to find *all* the cups in a given situation, or just *some* of them—or just a specific number of them—or *most* of the cups in a given situation. Similarly, we might want all the cups in a given situation to be clean, or most of them to be clean, and so on.

If a goal individual is associated with plural cardinality, I will call it **quantified**. Quantified goal individuals all link to a type in the RPC medium: in the case of 'all cups', the type would be 'cup'. They can additionally have a link to a property in this medium, so the agent can express a goal like 'all the cups are clean', or to a location, so the agent can express a goal like 'all the cups are in the cupboard'. Quantified goal individuals are also defined for one other feature, which is a quantifier. Quantifiers are either 'cardinal' or 'proportional': in my scheme, a cardinal quantifier is an absolute number (e.g. 'five') associated with the numerosity of objects represented in the salient region, or a relative change in numerosity, read from a change in the size of the salient region when the object's goal properties are used as a memory cue. Both these things are recorded in another field of the WM individual medium, which is tonically active, just like the other fields, allowing for quantifying determiners to appear at any head position in a DP when a goal individual is reported in language.

Note that a quantified goal individual can also be given other quantifiers. In particular, we can specify that *no* objects of a given type should have a certain property. For instance, we might specify that no cups should be dirty.

A token goal individual is inhibited when a single matching object is identified. A *quantified* goal individual is only inhibited when the right number, or proportion of matching objects are identified—as specified by its quantifier.

Consider a case where the agent wants to put all the cups (in the current environment) in a cupboard. I suggest a quantified goal individual is used to achieve this. This goal individual, when active, requires the agent to establish the set of cups in the current situation. This might be possible attentionally, if these cups happen to form a group, but it might require several separate actions of attention. In the latter case, the set of cups has to be established through a query to memory. A quantified goal individual prompts actions of attention, but also a query to memory.

8.6.6.1 Quantified episodes

A procedural alternative to a quantified goal individual would be a quantified episode associated with the given situation: in the current case, for instance, 'Put all the cups (/any cup) in the cupboard'. I like the idea that this can be active as well—but I think it's different: a quantified goal individual identifies a goal *state*, whereas an episode just provides an instruction. The state is: 'All the cups *should be* in the cupboard'.

Actually, I'm not sure how a quantified instruction like 'Put all the cups in the cupboard' would work. The quantified *assertion* 'Ali put all the cups in the cupboard' is a stative assertion in my model: it establishes all the cups in the relevant situation, and then asserts that the query 'Ali put X in the cupboard' retains all of these cups. So how do I *execute* this quantified instruction?

An earlier proposal I had was that there are two separate mechanisms: an attentional mechanism, identifying the set of cups and selecting them one at a time; and a motor mechanism operating on each selected cup and executing an action on it. That's actually not too different from what the quantified goal state achieves. (Except it's not important *who* achieves it, or *what actions* achieve it.) In either case, the termination conditions are important: when do you know the goal is achieved?

8.6.6.2 Reactivating an achieved goal

Say I put all the cups in the cupboard: I've achieved my goal, so the associated goal LTM individual can be inhibited. (I don't need to be looking around for cups any more.) Now during perception say I happen to notice another cup. Does this serve to *re-activate* the inhibited goal individual? My guess is that since a quantified goal individual describes a general constraint about a situation, it's a constraint that's renewed in time, perhaps gradually. So after having finished clearing up the cups, if a new cup arrives, I might not *immediately* clear it up: but after some time passes, I will gradually restore the goal of clearing up the cups, and then clear up the new cup.[11]

8.6.7 Learning associations between contexts and goal individuals

In Figure 8.3, there's a function (f) that maps contexts (LTM environments, LTM times, LTM situations and goal environments) onto goal individuals. How is this function learned?

It's well known that individual objects can be associated with rewards, and that these associations are context-dependent (see e.g. ??). It' also well known that agents actively seek objects associated with rewards (see e.g.). My basic assumption is that the set of goal individuals active in a given context identify those objects that are associated with reward in this context, and that therefore influence the agent's behaviour. Based on this assumption, it follows that the function f is trained on occasions where an agent receives a reward in the presence of an actual object.

In the model I propose, f is trained when the agent receives a reward *when attending* to an actual object. If the reward is due to the object, the agent typically has to attend to the object in order to receive it. For instance, say the agent takes a bite of a cake, and gets an (unconditioned) reward from the taste: at the moment this reward arrives, his attention will be on the cake.[12] It's in this situation that the agent generates training data

[11] This has somewhat the flavour of Mayr and Keele's (2000) 'backwards inhibition' phenomenon, in which subjects are slower to adopt a task if it is one they have recently been engaged in.

[12] Or perhaps the taste of the cake *draws* his attention to the cake. There's more about this in Chapter 11.

for the function that maps from his current circumstances to a goal individual associated with the cake.

The specific learning mechanism I envisage is exactly the same as the one proposed in the model of motor learning in Lee-Hand and Knott (2015). In that model, basic simple reach and grasp actions are learned through intrinsic reinforcement signals associated with touch sensations on the hand. Here, the infant has to learn a function mapping a visual representation of a target object onto a goal motor state, associated with a successful reach or grasp action. During training, an infant is presented with a target object, which he attends to, to create a visual representation of the object; he then executes a hand/arm action at random. In some cases this action happens to result in his arm touching the target, which generates an intrinsically pleasurable tactile sensation. In this case, a piece of training data for the function can be identified.[13] The training item is a tuple mapping the infant's current visual representation of the target object onto the current motor state. However, note that the function to be learned needs to deliver a *goal* motor state, not an actual motor state. In Lee-Hand and Knott's model, the training item is created by *copying* the agent's current motor state at the time of the tactile reward to a parallel medium holding goal motor states, and then storing an association between the visual representation and this newly-created goal motor state. Axiomatically, at the time a reward is received, the current motor state is a goal motor state. After training, when the infant is presented with an object, the learned function maps the visual representation of this object onto a goal motor state, which is different from the current motor state; a motor controller takes these two states, and delivers a motor impulse which moves the current state in the direction of the goal state.

Analogously, with this mechanism, I assume that if the agent is attending to an actual individual, and receives a reward, he logs a piece of training data for the function f, that maps his current circumstances onto a goal individual. Specifically, the properties of the actual individual he is attending to are copied into a parallel medium holding the properties of goal individuals (the 'goal WM individual' medium), and these properties are associated with a goal LTM individual. I assume that the activated goal LTM individual can either be an existing one, if there is already a goal individual associated with these properties, or a new one if not. (That is, I suggest that there's a mechanism for 'recognising' goal individuals, that's entirely parallel to the mechanism for recognising actual LTM individuals, that was discussed in Section **??**.) During training, that probably happens offline, the activated goal LTM individual is mapped to the activated set of goal properties, and the current context (LTM environment, LTM situation, LTM time, goal LTM environment) is mapped to the activated goal LTM individual.

[13] Training doesn't actually occur at this time: all that happens is that a training item is *remembered*. Training happens offline, on a large collection of stored training items.

8.6.8 Goal individuals and language

I want to suggest that stative sentences featuring a modal auxiliary like *must* or *should* describe the properties of goal individuals. For instance 'A window should be open' describes the property of a goal individual.

Goal individuals also feature in sentences featuring modal verbs like *seek* or *find*. (In the former case, the goal individual is asserted; in the latter it is presupposed.)

My main idea is that modal auxiliaries, or modal verbs, signal the entering of a special cognitive mode called **goal reporting mode**, in which information in the goal WM individual medium is simply copied to the (actual) WM individual medium. We do not want goal WM individuals to have their own interface with language, because if they did, there would be no reason to expect the words denoting goal objects to be the same as those denoting actual objects.

To illustrate how goal reporting mode works, consider how the agent would execute the action of *seeking a cup*, and then report this action. John first attends to himself. Then he activates the goal individual 'cup', sufficiently strongly to put him into task-execution mode. This actually generates a particular category of motor action, which is an action of *seeking*, that involves directing attention here and there, moving around, lifting things, and so on. However, these actions are not normally reported, since recording only happens when an episode is completed, and this requires that the sought-for object be found, as discussed in Section **??**. At the same time, we do not want John to suspend his ability to *record what he's doing* in the interim. He must be able to record what he's doing at any given time. If queried, he must be able to say *I'm looking for a cup.* Also, at the end of a search, whether successful or unsuccessful, he must also be able to record his search activity as part of episodic LTM, particularly if it took a certain amount of time or effort. In either case, I suggest that he activates goal-reporting mode, to copy the 'cup' goal WM individual to the actual WM individual medium, and then copies this representation to the 'patient' medium of the WM episode. Now he has a complete WM episode, since it includes a patient, and an action ('seek').

Want is another interesting action. Here again, it's a meta-level action, whose only effect is to enable a particular mode. But in this case, there's no first-order activity to report: what's reported is simply the existence of a goal individual.[14]

8.7 Actions of creation

In this section I'll outline a model of actions that create new objects. This could include very simple actions, like drawing a square, or more complex actions, like making a cup of tea, or extremely abstract actions, like getting a degree.

In each case, I'll argue that the object to be created is initially represented as a *goal*

[14] *Wanting to* is also interesting: this reports a desired action, rather than a desired individual. A model of this needs to reference the fact that the desired episode doesn't have to be immediate: it could take place in the future. [I'm halfway through this thought. . .]

individual, of the kind described in Section 8.6. I suggest that there is a special meta-level action that an agent can perform 'on' a goal individual, which is to create it. The result of this action, eventually, is the appearance of the actual individual. An action of creation is ostensibly a transitive action—the verb that describes it takes a transitive object. I want to assimilate the sensorimotor routine associated with actions of creation to the scheme I have already developed for simpler transitive actions like reaching-to-grasp, as described in Knott (2012). In my model of reaching-to-grasp, the target object to be grasped is represented twice, in different modalities, at different times: to begin with as a visual representation, that maps to a goal motor state, and at the end of the action, as a current motor state and a haptic stable-grasp signal. I argue that the object of an action of creation is likewise represented twice in different modalities: at ths start of the action it is represented as a selected goal individual, and in the consequent state of the action it is represented as an actual individual.

Recall from Section 8.6 that a goal individual is linked to a structure of goal properties. A key assumption in the model of creation I present in this section is that a goal individual can also be linked to a *LTM situation*, of the kind described in Section **??**. Recall from that section that a LTM situation holds a quasi-spatial representation of the structure of a task: the currently active LTM situation provides a tonic bias on the recurrent SOM that computes a probability distribution over episodes at any given time, and modifies this distribution each time an episode is experienced. This tonic bias essentially configures the recurrent SOM to define a particular set of potential episode sequences, which can be thought of as 'ways' of achieving a given task. The LTM situation associated with a goal individual is one which defines *the task of making an instance of that individual*. The main proposal in the model of creation actions is that the meta-level action of 'making', when executed on a given goal individual, causes the agent to activate the LTM situation associated with this goal individual—that is, to *adopt* the task of making an instance of this individual. Of course, this new LTM situation not only defines a particular set of possible episode sequences, which are 'ways' of making the individual—it also induces a new distribution over *other* goal individuals, which are the component objects from which the individual is to be assembled, and the tools that will be required. I will discuss how episode sequences and goal individuals interact within situations in Section 8.9.

8.7.1 Termination of an action of creation

In an action of creation, there may be several episodes. After each of these episodes, the 'current context' of the recurrent SOM structure is updated, while the LTM situation that represents the complete creation plan remains tonically active. However, when the final episode is executed, and the object is completed, there must be a mechanism that inhibits the currently-active LTM situation, and activates another situation. There are two questions: firstly, how does the agent recognise that the current plan has been completed? Secondly, how is the next situation determined?

As to how the first question, I suggest that the goal individual associated with the LTM situation serves to represent the state

As to the question of which situation is activated next, there are a few alternatives. Perhaps the next situation is the situation in which the just-completed situation was embedded. Perhaps the agent is in the middle of several ongoing situations, and has to choose one of these to return to.

8.8 Situations and the creation of episodic memories

8.9 Interactions between episodes and goal individuals

In the model introduced above, when an agent embarks upon making a cup of tea—that is, when he 'enters' the LTM situation associated with the goal individual 'a cup of tea', there are two parallel mechanisms that help to guide him through this task.

One mechanism is implemented in the circuit involving recurrent context-updating SOM. (Recall that this SOM takes input from the currently active LTM situation, as well as from its recurrently updated context representation.) This circuit begins by delivering a distribution of possible episodes that the agent can execute as a first step in making a cup of tea. If one of these episodes is selected, it then delivers a distribution over episodes that should happen next, and so on. If the agent makes tea habitual way, for instance beginning by filling the kettle and proceeding by putting tea into the teapot, there are particular sequences of episodes that this circuit can learn.

The other mechanism is implemented in the circuit involving goal LTM individuals. The tea-making LTM situation also induces a distribution over goal individuals: namely the individuals which the agent should search for, and act on, in order to produce a cup of tea. These include a cup, a kettle, a teapot, some tea, some water, and possibly other things. As already discussed in Section 8.6, goal individuals can have properties, or even hierarchical structure, so that they support not only the *finding* of the individuals that are required, but also actions that change the properties of the found individuals when necessary, so they are appropriate for the goal. For instance, as described in Section 8.6.5, the goal individual representing a kettle in a tea-making scenario can be specified as containing water; processing this individual can generate the preconditions for an action of transferring some actual water into the kettle.

In the complete scenario I envisage, the distribution of possible next episodes generated by 'habit' is supplemented with a distribution of actions activated by the processing of goal individuals. Sometimes, the action selected by the agent is selected because of its position in a familiar routine; in other cases, the action is selected because the agent is explicitly processing stative goals associated with the current task, and selecting actions that achieve these goals. Perhaps the agent can move between these modes somewhat consciously, but I like the idea that they can operate simultaneously as well.

In some cases, perhaps *all* the actions required to create a given goal individual can be derived from a representation of the goal individual. For instance, a cup of tea is a cup, with milk and tea in it. The tea is hot water, with tea infused in it. However, quite often the history of steps involved in a creation process cannot be derived from the way the final product is represented. (I think tea is a case in point.) The recurrent SOM is therefore sometimes essential, not just in representing highly compiled habits, but also in representing the sequential steps of learned processes, of the kind that can be communicated in instructions or recipes.

8.10 LTM situations

8.10.1 Types of situation

Situations can be can be triggered by different types of event. In this section I will review the different circumstances which can lead to the activation of a new situation.

8.10.1.1 Situations associated with acts of creation

For situations associated with acts of creation, there's a fairly obvious attribute, which is the goal individual to be created. If I decide to make a cup of tea, the situation I will activate is the one associated with the goal individual 'a cup of tea'. If there is no associated situation, I will create one, and activate that—and this newly activated situation unit will record any successes in my exploratory attempts to make a cup of tea, whether these take the form of episode sequences or helpful intermediate goal individuals, to guide any future occasions when I embark on making a cup of tea.

A wider question is when goal individuals are activated. Sometimes they are simply activated as a function of the agent's internal drives: this case will be considered in Section 8.10.1.2. Sometimes they are activated as part of a wider situation: this case will be considered in Section 8.10.1.3.

8.10.1.2 Situations associated with drives

Some goal individuals can become active simply as a result of the agent's internal drives. For instance, the agent might develop a desire for a cup of tea: in the current model, this would be implemented simply by the activation of the goal LTM individual 'cup of tea'. This models the scenario where a cup of tea has become a conditioned stimulus for the agent, through having become associated with an unconditioned reward. These associations can be context-specific, as already discussed in Section 8.6.7.

The termination conditions for a situation associated with a goal individual are also obvious: the situation terminates when an actual individual is found that matches the goal individual.

8.10.1.3 Situations activated as components of larger plans

One of the elements of my weekday morning routine is to make a cup of tea. This happens right after I wake up.

As already discussed, this is a plan in its own right, and is represented by its own LTM situation. When I'm making my cup of tea in the morning, I'm in the middle of a larger plan. When I've finished making the tea, and terminate this plan, I will reconsider the plans that are associated with the current moment, and will thus be able to resume this larger plan. (If there are any other ongoing plans, these will also compete to be selected.)

8.10.1.4 Situations associated with times

Some situations are activated by times. For instance, on the morning of a normal weekday, I wake up, make tea, have a shower, have breakfast, and walk to work. In my model, this plan is implemented in a situation, which conditions a particular sequence of episodes.[15] This plan is activated by the fact that the current time 'now' is a particular the time of day, and a particular time of the week. I presume the situation is terminated when the steps of the plan are achieved, or when the time unit allocated for it is elapsed. In the former case, we allow the possibility that the situation continues beyond the time period allocated for it, and the agent is 'late' in finishing the plan. In the latter, we allow the possibility that the agent gives up halfway through a plan.

If I had to do something different on one particular morning—for instance, catch a plane—a token situation would compete against the habitual generic situation. We have to envisage that several situations can compete for selection, with only one being picked at any given moment. Of course, as discussed in Section **??**, there is a provision for the current situation to be suspended, so that other situations also associated with the current time can be activated.

8.10.1.5 Situations associated with unexpected episodes

I suggest that situations can also be associated with specific *episodes*. For example, say I'm walking through the park, on my way home. I assume I'm in the middle of a 'walk home' plan, where the desired state is one where a particular goal individual who happens to be myself is at the location 'home'. Since I'm in this situation, my actions are conditioned by the goal of getting home. But at the same time, as discussed in Section 8.6, I can still perceive episodes unrelated to this goal, if they are salient enough. Say I observe an episode in which a stranger jumps out at me from behind a tree. This should have the effect of activating a *new* LTM situation, unrelated to the situation of walking through a park. A normal episode perceived while walking through the park will cause an update in the recurrent context-representing SOM. But in this case, I switch much more discretely to

[15] It also conditions a particular set of goal states: I should be caffeinated, clean, full, and at work. If my regular sequential routine is disrupted, these will ensure my actions will have the same effects as the habitual sequence of actions.

a new situation, which is a function of the current WM episode *by itself*—and, I presume, the recurrent SOM's representation of 'the current context' is initialised to some standard value, representing 'start of the situation'.

Braver and Cohen's (2000) model of how plans are learned provides a very useful reference point here. What is relevant about the stranger jumping out at me is that it is a good *predictor of reward*—or in this case, a predictor of punishment.[16] If I don't take action, there is a good chance that I'll be attacked, mugged, beaten up, all of which are punishing stimuli. Braver and Cohen's model is a model of how an agent learns to interrupt a current plan and adopt a new one. In their model, a plan is an assembly in prefrontal cortex that biases behaviour in a particular way. But when the plan is first adopted, the agent doesn't know how his behaviour should be biased: the selected prefrontal assembly is activated simply as a function of the current perceptual stimulus. The only significant thing about this stimulus is that it is a predictor of reward or punishment. However, any exploratory behaviour that occurs at this point that proves useful in obtaining the reward, or avoiding the punishment, will become associated with the newly active assembly, so that next time the reward-predicting perceptual stimulus is encountered, it will trigger the establishment of a specific cognitive set, encoding behaviours relevant to the situation.

In my model, I envisage a similar kind of learning takes place. A separate mechanism, not shown in Figure 8.3, learns to activate predicted rewards, or punishments, associated with perceived episodes. If the predicted reward/punishment exceeds some threshold in magnitude, the episode is mapped to an associated LTM situation. (In my model, LTM situations play the role of Braver and Cohen's PFC assemblies, and WM episodes play the role of 'perceptual stimuli'.)

There's an interesting connection to language in this model. When we report a plan-changing episode in a narrative, we use the word 'when', in a special sense:

(8.1) I was walking through the park on my way home *when* a man jumped out at me from behind a tree.

In the discussion of *when* in Section **??**, I suggested it was used in a context where the episode it introduces functions as a cue to episodic LTM. In the current case, *when* does not function to establish a distant context in LTM: it simply introduces the next thing that happens in the narrative. However, I suggest that the episode still functions by itself as a retrieval cue: not to retrieve a temporally distant *context*, but to retrieve a particular *LTM situation*.

I think the situation associated with a reward-predicting episode terminates when the reward is achieved (for positive predicted rewards) and when the reward is prevented (for negative predicted rewards, i.e. punishments). Therein lies the basis for stories. (As usual there's also a possibility that the situation gets abandoned, if something else becomes more important.) My guess here is that the termination condition has something to do with reward. If the predicted reward is negative, the purpose of the situation is to

[16] Episodes can also be predictors of positive reward: for instance, if I'm walking through the park and I see a great photo opportunity, I might be inspired to interrupt my walk and take photo.

generate actions that avoid this anticipated negative reward. If it is avoided, there will be a dopamine burst (see e.g. Schultz *et al.*, 1997): I think this might be the trigger to terminate the situation. If the predicted reward is positive, the situation is designed to generate actions that ensure it happens. If it doesn't happen, there will be a dopamine drop (see again Schultz *et al.*), so again the moment of predicted reward is the moment at which the situation can terminate.[17]

8.10.1.6 Planned token situations

There's interesting recent evidence that imagining or planning episodes in the future involves the same mechanisms as are involved in representing episodes in the past (see again Hassabis *et al.*, 2007; Martin *et al.*, 2011). In the light of these findings, I propose that when an agent plans something for the future, he creates a new token instance of a LTM situation, and associates it with a future time. Or configures it so that an anticipated future episode will trigger activation of this situation. A key thing about a plan is that it can have internal structure: it can involve several episodes, that occur with a predefined sequential structure. A situation is a good way of implementing a plan.

When the agent activates the relevant future time period or future episode, the planned token situation competes to become active, in a way that's similar to situations that routinely become active at such times, or following such episodes. If the situation competes strongly enough, it will be 'entered', and the agent's subsequent behaviour will be determined by the associated plan.

Again, a context is a little like a spatial environment. Within a context, there can be multiple episodes that compete to control experience: these are like multiple places in the current environment that the agent can get to. But there can also be multiple situations that compete to be activated: these are like places that can be reconstrued as whole environments, with their own internal structure. x

8.10.2 How LTM situations are learned

A question that hasn't yet been addressed is how an agent first learns LTM situation representations. Since an LTM situation is a localist assembly, the first time it is activated, it can't yet have any meaningful associations. Under what circumstances would a new LTM situation be activated?

I suggest the basic mechanism involves a process whereby LTM situations are *recognised*. Sometimes, an agent should realise he is in a situation of a kind he has been in before, or perhaps in exactly the same situation he has been in before: this is analogous to recognising the type of an object, or a token object. At other times, he should realise he is in a *new* situation, that he has never been in. He doesn't yet know what to do in this situation, but he can still activate a new assembly that will hold the learning that will take place in it, now and in the future.

[17] I presume it's not just *any* reward, but a rewarding episode, with some content to it: for instance, getting beaten up (for a negative reward) or getting a good photo (for a positive reward).

I suggest that creating a new LTM situation unit is similar to creating a new LTM individual unit: it is something you do when you encounter an individual, or a situation, that you don't *don't recognise*.

In Section 8.10.1 I enumerated the different triggers that can become associated with LTM situations. My basic idea is that if these triggers fire and there is *no* associated LTM situation, then a new one is simply created. For instance, if an LTM goal individual is selected and there's no associated LTM situation, one is newly created, and is used to hold any learning that subsequently happens about how to make or find this goal individual; similarly, if an episode occurs that predicts an unexpected reward and there's no associated LTM situation, one is newly created.

8.11 Situations representing joint actions/social scenarios

The situation medium might also be able to represent culturally conventional scenarios involving multiple participants playing particular roles. I'll take the example of a restaurant meal. In this scenario, there's a waiter and a group of diners, and various actions that happen in a structured way: for instance, the diners enter the restaurant, the waiter greets them, finds a table for them, they sit down, the waiter brings them menus, they choose food, the waiter asks them what they want, and so on.

8.11.1 Preliminaries: representations of social roles

Roles are object types, in some sense: you can look at John and say 'John is a waiter', in the same way as you can say 'John is a man'. However, roles are defined in large part because of the actions that someone does. (There are certainly physical properties too, like wearing an apron, but these are secondary: they might allow a waiter to be identified, but they're not integral to the meaning of 'waiter'.)

I envisage a model where an observer learns roles as frequently co-occurring propensities to participate in particular episodes. For regular object types, the things that frequently co-occur are literally present simultaneously to the senses: they are low-level perceptual properties, activated in parallel in the RPC. For roles, the things that co-occur are remembered dispositions associated with individuals. I assume we learn about roles by watching individuals over extended periods of time, and recording stative properties about these individuals that relate to their propensity to participate in particular types of episode. I assume these are stored in associations between LTM individuals and abstracted LTM episodes in the c-EP SOM, of the same type that feature in relative clauses. So, for instance, when we observe John, we might notice that he often greets people, gives people menus, asks what they want to eat, brings them food. We record these facts as properties of John: that is, pointers from the LTM individual 'John' to abstracted LTM episodes like 'X greets customer', 'X gives customer a menu', 'X asks customer what they want to eat' etc. Recorded facts of this kind provide the *input* to the system that learns about social roles like ‘waiter’, ‘doctor’ and so on. This system learns about *correlations* between episode-based properties, in the way that the system learning concrete object types learns about correlations between perceptual features. (The difference is that the system learning social roles operates on representations of *LTM* individuals, whose properties generalise over time, while the system learning concrete object types operates on *perceptual* representations of individuals.)

8.11.2 Preliminaries: initiating a social scenario

An observer is able to recognise that he is ‘in’ a social scenario, and in some cases to *initiate* a social scenario. This involves several things, some of which relate to temporal aspects of ‘the current context’ representation, and others to spatial aspects.

On the temporal side, a social scenario provides an input to the situation SOM, that biases its predictions about the next episode. When a scenario is started, we should reset the ‘previous situation’ representation to a default value, to indicate we’re at the start of the scenario.

On the spatial side, certain scenarios are associated with certain physical spaces: for instance, the ‘restaurant meal’ scenario typically happens in a physical restaurant. (The ‘lesson’ scenario typically happens in a classroom.) But equally importantly, within a given instance of a scenario, the roles (like ‘waiter’, and ‘diner’) have to be *bound* to particular token individuals. Entering a scenario requires these binding operations to happen. There needs to be a *mechanism* which *systematically* finds bindings for all the relevant roles. I’ll talk about that in the rest of this section.

I suggest that roles can also be thought of in a quasi-spatial way, as places that token individuals can occupy. A scenario can then be thought of in a quasi-spatial way, as an environment, within which these role-places are found. Role-places are different from places in physical space, in the sense that there can be no places ‘in between’ the places occupied by individuals, and there’s a requirement that an individal be found to occupy each space. This latter requirement is a way of expressing the fact that all the roles associated with a scenario have to be bound to individuals. I suggest the mechanism that systematically finds bindings for roles (mentioned above) makes reference to this quasi-spatial represenation of roles. When an observer enters a physical environment, he is obliged to scan it in some fairly exhaustive way to check for salient contents. I suggest that when an observer enters a scenario, and activates a quasi-spatial representation of the scenario as an environment, he is likewise obliged to scan the ‘locations’ in the environment (i.e. roles) in a systematic way, by activating each role in turn. When activating a role, an individual must be found to bind to the role.

I also suggest that these role-places are involved in an account of some kinds of definiteness, or uniqueness. A country is a social institution, in which there is one president (which can be different at different times). A race is a social institution in which there is one winner. When we say *John was the winner of the race*, we have a special kind of equative copula which is invertible (*The winner of the race was John*), but the equation isn’t a matter of identifying two individuals: ‘the winner’ is in some sense a predicate, but it’s a predicate which is constrained to be predicated of exactly one individual. Role-places can also feature in an account of definites in possessive constructions: I can say *The mother of my friend* because everyone has a mother: for each individual there is a unique individual that stands to them in the role 'mother'.

The binding of a role to an individual can happen in different ways. When assigning roles to external individuals, the process is sometimes a *perceptual* one, where the process amounts to *identifying* the individuals playing particular roles. There are two perceptual mechanisms that mediate this process. Firstly, recall that roles are associated with physical object properties: for instance, waiters have aprons, carry teatowels etc. So the role 'waiter', when activated, triggers a simple *visual search* for an individual with the right properties. If this search returns an individual, the role can be bound to this individual. Secondly, recall that scenarios are also associated with physical environments: in some cases, roles can be associated with particular *locations* in the physical environment. For instance, waiters are often standing behind the bar, while diners never are. (In classes, teachers are often standing at the front of the classroom, near the blackboard, while students are more typically seated at desks.)[18]

However, the binding of roles to individuals can also happen as part of the *creation* of a scenario, rather than as part of its passive observation. An agent can choose to play a role in a scenario. This can happen on different timescales. On a short timescale, an agent can choose to have a restaurant meal, and more specifically, can choose to walk into a particular restaurant, and thus bind himself to the 'diner' role for this particular instance of the 'restaurant meal' scenario. On a long timescale, an agent can choose a *profession*, like waiter or teacher, which dictates a regular pattern of activities, involving many specific 'restaurant meal' or 'lesson' scenarios. Finally, an agent who wants to initiate a particular multi-agent scenario can also make ouvertures to individuals, asking them if they would like to adopt particular roles. For instance, a waiter might ask passers-by if they would like to become diners in his restaurant. Scenarios involve co-operation between agents, and often require some negotiation to be initiated.

In all these latter cases, the binding of roles to individuals is a *decision*: a matter of *personal choice* for the individuals in question. Ultimately an agent will adopt a role because it's associated with reinforcement. For instance, an individual might choose to become a diner in a restaurant because it will likely be enjoyable. (Note his participation in the 'restaurant meal' scenario as a whole has to be represented as enjoyable, not just the individual episodes that take place while he's playing this role.) An individual might choose to become a waiter because it brings remuneration in a way that's more pleasant than other alternatives. All of these things have to be learned through reinforcement. Notice the activities in scenarios are rewarding for all participants: they're not a zero-sum game.[19]

[18] If you know you're in a cafe, you will actively seek out referents for the roles 'waiter' and 'customer'. But it's also possible to activate a scenario 'bottom-up', triggered by your classification of an individual in your current environment as a waiter.

[19] Even a slave doing work for a master is doing this work 'out of personal choice' in some sense, for reasons related to reinforcement/punishment: he takes on the role of 'worker' because it conveys the most

8.11.3 Where are social role-bindings stored?

After a token individual is bound to some social role like 'waiter' or 'diner', there has to be some memory of these bindings, so they can be consistently applied through the course of the scenario. What mechanism implements this?

First, consider what's involved in 'applying' bindings. Consider a case where I decide to adopt the role of waiter in a restaurant. The scenario's input to the sitations SOM biases it towards predicting the generic episode 'waiter gives diner a menu'. The generic episode makes no reference to me: it just refers to the roles 'waiter' and 'diner'. But *if I'm the waiter*, this generic episode should have the effect of making *me* give the diner a menu. Thinking sequentially, when I 'attend to the agent', my top-down expectation that this agent is 'the waiter' should have the effect of putting me into action execution mode. If I'm the diner, the very same expectation should have the effect of putting me into action perception mode, and triggering a visual search for a third-party individual with the kind of perceptual properties associated with waiters.

I suggest these effects are achieved due to representations stored in the *WM individuals system.* An observer stores the properties of recently-attended individuals in WM individuals, which are convergence zone units that hold associations between location, number and perceptual properties in short-term weights. This medium can also be used to store his intention to participate in a social scenario in a certain role, or the fact that some other individual is participating in a scenario in a certain role. (...)

Actually, I think role-bindings might also have to be stored in the episodic *LTM* system. This is because scenarios can be interrupted for arbitrarily long periods of time, and their role-bindings must be remembered over these periods. Perhaps the bindings for the *currently active* scenario are stored in the WM individuals system—but when a scenario is interrupted, these bindings have to be stored in episodic LTM. (*I think they're stored in a LTM episode, in fact, but I haven't written about this yet.)

8.11.4 Scenarios in hierarchical LTM representations

Since a scenario is a kind of task (what I earlier called an 'LTM situation'), it's also implicated in the formation of hierarchical structures in LTM. I can represent the internal structure of a restaurant meal, describing the actions of the diners and waiter and their sequential structure, but I can also refer to the whole meal as a single episode, in which 'the diners *go for* a meal', or 'the waiter *serves* the diners a meal'. The episode can be construed from the perspective of the diners, as an action of theirs, or from the perspective of the waiter, as an action of his. In other examples, the participants are equal: for instance, the participants in a game 'have' a game, or 'play' a game *collectively*; the competitors in a race 'have', or 'run', a race *collectively.* The participants of a fight 'have' a fight, or simply 'fight'. (The fight or the game each involve several actions, with some sequential structure.)

benefit to him, within his limited range of options.

I think the way these scenarios are represented as wholes is as *individuals*. That's certainly what's conveyed linguistically. There are special verbs like *have* and *do* and *play* that express the action of an individual's participating in a scenario, represented as patient. And there are special nouns that represent scenarios as individuals, like *meal* and *game* and *race*.

If scenarios are construed as individuals (i.e. objects), there are some interesting ways of exploiting the universal duality between objects and environments. Recall that every object can be established as an environment. What's the environment associated with a scenario-represented-as-object? This would be a 'place' in which the 'parts' of the scenario are found. When we talk about a table as an environment, we represent locations within it, for instance the top, or the edge, which can contain other objects (e.g. books). When we construe a scenario as an environment, I suggest we get the quasi-spatial representation mentioned above, in which the locations within it are the roles that it involves. These locations can be 'filled' by individuals, in the special sense described above. This analysis also explains why you can use possessive constructions to refer to roles: as in 'the winner *of* the race/game', or 'the teacher *of* the class'.[20]

An agent who witnesses a scenario, either as participant or observer, can store it in LTM at two levels of detail: either a coarse level ('We had a meal at the Esplanade') or a fine level ('We had a meal at the Esplanade. We went in; the waiter showed us to a table by the window; he brought us menus; we ordered pizzas; we waited; the waiter brought us the pizzas; we ate them; we paid; we left'). We can also choose to focus on the distinctive things: 'We had a meal at the Esplanade. We ate pizzas.' I need a nice model of how these hierarchical structures are stored. I'll make a suggestion in the next section.

8.11.5 Scenarios and hierarchical representations in episodic LTM

I can say:

(8.2) We ran a race. Then we went for a meal. Then we went home to bed.

Or I can go into some detail:

(8.3) We ran a race. [We lined up at the start, waiting for the starting gun. At the gun, we took off. I crossed the finish line first, Bob was second.] After the race, we went for a meal at the Esplanade. [We walked in and sat down, and the waiter took our order. Bob had chicken, I had fish. We paid the bill and left the restaurant. Then we went home to bed.]

My questions are:

- How are the coarser-grained episodes represented (i) during experience? and (ii) in LTM?

[20] This doesn't work very reliably: you can't say 'the waiter/diner *of* the meal'. In this case the participants are more easily thought of as belonging to the physical place where the scenario occurs—which of course is another equally good way of locating them.

- How is the *transition* between coarse-grained and fine-grained episodes effected during recall?

How are the coarser-grained episodes represented during experience and in LTM? A coarse-grained episode like *Bob and I ran a race* can't be represented all at once during experience, because it takes a while to happen. The WM episodes that make it up are much smaller things. The coarse-grained episode has to be *constructed*, from meta-level operations that relate to the initiation of a *LTM scenario*, and its subsequent termination. Crucially, the meta-level operations are done by the participants of the scenario, so they are things that can be experienced.

The initiation and termination actions are a bit different. The initiation one is something like *Ali decided to run a race*. The termination one is something like *Ali finished the race*, I think. Importantly, the expression *Ali ran a race* implies both the initiation of the scenario, and its natural termination: 'the race' describes an event, but this event does not fully exist until it's also terminated. Then it's an episode with a certain status in LTM.

I think when I *initiate* the race scenario, the *whole scenario* has to be encoded into episodic LTM. It's still a plan—but we now know that plans are stored in epsiodic/hippocampal LTM, as well as memories of past events. As I monitor the individual events making up the race, I progressively move parts of the race scenario from 'the future' to 'the past'. This difference coincides with a difference in modality: the things in the future are probability distributions, because we don't know which of them will happen; the things in the past are single episodes, because exactly one episode out of the many possible ones occurred at every stage. (Though we are also free to explore counterfactual alternatives.)

Aside: representations of time and modality in episodic LTM I think to come up with a good story about hierarchically structured representations in episodic LTM, I have to think a little more about how time and modality are represented in this medium, and then return to the question of hierarchy. I'll do that in this section.

My earlier idea was that there's a representation of a special time, 'now', which points to the 'current situation' related to each scenario. Whenever the situation is updated, by the occurrence of a new episode, this pointer is re-assigned. I assume the pointer is implemented in long-term weights, rather than short-term ones, so that a scenario can be interrupted for a long period of time. This means that re-assignment must involve setting one long-term association, and also *actively resetting another long-term association to zero*.

When you activate 'now', you need to retrieve a finite set of situations-within-scenarios.[21] In practice, 'now' could just *point to a set of units in the situations SOM*. From these, we can reconstruct a set of candidate scenarios that I'm in the middle of: having a meal, getting a degree, doing an assignment. Then I can choose which of these is most important, and activate that one. I'll also start it at the right place: the 'now' place.

[21] Before you can retrieve a set of situations, you have to select *a particular agent* and adopt his/her perspective, as I describe later on. This involves activating a WM episode with the agent as Agent, and the Action set to 'do'.

Say I restart 'having a meal'. I'm in a situation where I must next tell a waiter what I want to eat. I do this: now I'm ready to update the situation SOM. Crucially, there's another step in this process: before I do the update, I *zero* the LTM association between 'now' and the current situation. Then after I do the update, I *potentiate* the LTM association between 'now' and the *new* current situation. If I then switch to another task, I'll know where I got to. I can get this by activating 'now', or more precisely, by activating 'now' *and* the scenario S, whatever it was.[22]

I assume I can also plan for things to happen at specific times in the future using a variant on the above scheme: associating scenarios with *future* times, like 'next Monday'. Something like this must be possible, if things in the future are also in episodic LTM. But in this situation, how can I distinguish future episodes/scenarios from past ones? It must be possible to query LTM for a past episode. (Though I guess you can also be 'reminded' of future, planned situations.) The mechanism that does this has to allow for multiple tokens of any given scenario, that occur at different times and/or places.

Here's an idea: the situations associated with future episodes will always be under the scope of some modal operator, 'want' or 'fear' or something like that. So a situation associated with a desired/feared future episode E *will always be represented by a different unit in the SOM* than a situation associated with an experienced episode E.[23]

At the same time, scenarios must still be associated with times. I think you have to also activate a time before choosing a scenario. I suggest the activation of a time can happen either before or after activation of the special episode 'I want'. The time doesn't have to come under the scope of the modal. If you choose a time, and then *I want*, the chosen time will obviously affect the distribution of episodes/scenarios. If you choose *I want*, this will first affect the distribution of times, and you'll have to select one of these times: and then you have to choose an episode/scenario.

At this point, *you associate the chosen episode/scenario, and chosen time, with the active unit in the situation SOM.*[24] Now you're able to return to the current actual situation. At any point, to examine the plans you had for yourself *for the current moment*, you have to actively activate the 'I want' ('I decide') episode[25], and then use this as an LTM retrieval cue, *along with the current time* (represented absolutely, rather than as 'now').

One problem: on this model, if you've asserted that you want E to happen at some point in the future, and you try to query LTM for *past* occurrences of E, you'll still retrieve

[22] Again, the observer also has to establish a particular agent's mental context, so that the activated scenarios are those that *this agent* is 'doing': again, see below. (*What's John doing?* is different from *What's Mary doing?*.)

[23] When an episode is actually completed, it can be asserted into episodic LTM *without* any modal operator.

[24] Actually, I think if you do this, you have *decided* to do it rather than just recording your desire to do it. If you're just expressing desire for the episode/scenario, you're just reporting the emotional associations that this episode/scenario generates, at this time, in this mental state context. (You may be using your emotions 'empathetically', to) evoke *Bill's* feelings about the episode/scenario.)

[25] Or perhaps the 'I wanted/decided' episode. The states of wanting/fearing themselves should be associated with time periods: we can recall what we wanted or feared at a given time in the past, regardless of whether or not it happened.

the future E. Perhaps another thing to do is to add a 'subjunctive' flag to WM episodes, and set this to 'true' when representing the content of 'wanted' states. When we query the past, we can then set this flag to false (and ensure that the episodes we retrieve also have it set to false). If the subjunctive flag is part of the WM episode, then it can surface in the right ways in clause syntax: either as part of the verb conjugation paradigm, or as a separate particle (as in Russian, Arabic).

Back to hierarchical structures in LTM I now have a kind of answer as to how a scenario gets into LTM *when initiated*, and how to represent in LTM that it's not completed—in fact, how to represent each different stage of its completion. A special situation arises when the scenario is *fully* completed: for instance, when I finish the race I was running. The memory operation that encodes this is somehow implicated in a LTM representation of the past coarse-grained episode 'I ran a race'. How does a representation of the *completion* of a scenario enable it to be newly represented as a whole?

First, here's a concrete idea about how completion of a scenario is registered. Say a scenario has a special 'completion situation' in the situation SOM. This will be generated as a result of the situation-update operation following the last episode in the scenario (e.g. the diners walking out of the restaurant, or the medal ceremony of a race). In the completion situation, the participants of the scenario *relinguish* their participant roles; they no longer have any obligations to behave in a certain way, in accordance with the scenario.[26] So: in this completion situation, the episode can be represented as a whole.

One important thing that happens at this point is a second, higher-level *situation update operation*. We're now in the situation where the whole scenario is complete—so we can ask what happened *next*, just as we do when a complete episode updates the current situation. I assume it's the *same Situation SOM* that gets updated. But it has to be updated differently. When the situation is complete, I propose the following things happen. (1) First, the completion situation is *replaced by the situation in which the scenario was initiated.* (Which must somehow be stored in memory.) (2) Then the coarse-grained episode which represents the particpants' participation in the scenario is expressed *in the WM episode buffer.*[27] (Thus this buffer can hold episodes and meta-level (coarse-grained) episodes.)[28] (3) Finally, a *second* situation-update operation happens, to generate a representation of the situation obtaining as a consequence of the whole (completed) scenario. Importantly, this coarse-grained update operation *makes no reference to the active unit in the 'scenario' medium.* The scenario is represented *as an episode*, rather

[26] In some scenarios, the completion situation relates to the achievement of a goal, but I won't think about those for the moment. I imagine the completion situation is sometimes associated with a reward state of some sort (sometimes perhaps a 'social' reward, of the kind an agent gets from participating in a successfully completed scenario). Other times it might just relate to the point when the power to predict what happens next is lost: that also creates a natural boundary (of the kind discussed by Zacks in his account of event boundaries).

[27] Maybe operation (2) happens first: we use the active scenario to retrieve the desired episode into the WM episode buffer. We can use this as a retrieval cue to the situation SOM, to get back the situation in which the episode (qua *desired* episode) was initiated.

[28] I think it had to be represented in the WM episode buffer when the scenario was *initiated* as well.

than as an influence on the situation SOM generating the internal structure of episodes in the scenario. Importantly, the second time the episode is asserted, it's not nested within a 'desire' context. The content of the desire is now asserted *as fact*.[29]

Now consider what happens when we *retrieve* the scenario-episode from LTM. We have two options. (1) We can retrieve the *initation* of this episode. We do this by *activating the relevant unit in the scenario SOM*. What will happen now is the episodes *in* the scenario will be replayed, one by one, until they're all done (and then we'll deactivate the scenario and activate a representation of the consequent state of the whole scenario, as described above). (2) We can choose *not* to activate the relevant unit in the scenario SOM, and do a situation update operation without it: in which case we'll get to the consequent state.

Note these two operations correspond really nicely to the two points you can return to with a 'when' clause, as discussed by Moens and Steedman. One is the consequent state; the other is the initial state.

Following these ideas, we can propose specific mechanisms for the statements that *John started to [run a race / have a restaurant meal etc]* and *John finished [the race / the meal etc]*. The former is what actually happens in SM experience when John embarks on the scenario, and activates it as a unit in the scenarios medium. The latter is what happens when there's a double situation-update: the first one establishing that the scenario is complete; the second one updating to the situation that follows the scenario represented as an episode.

'Doing' The word *do* has the ability to refer to actions as objects, at a meta-level. (*What are you doing?* can receive an answer like *I'm drinking beer*, *I'm running a race.*) I think perhaps the meta-level actions referred to in the above discussion (e.g. activating a 'race' or 'cafe-meal' scenario) are implicated in an account of the verb *do*. For instance, the question *What are you doing?* can be answered by taking the 'current time' and seeing what situations are activated.

Attributing scenarios to individuals It's important to distinguish between what *I'm* doing and what *someone else* is doing. We already know that the situations medium can distinguish between my plans and someone else's plans. I presume that the scenarios medium *doesn't* distinguish between agents.[30] (That is, that the same 'cafe-meal' scenario serves to represent my cafe meals and someone else's.)

Note that when I *initiate* a plan, it's nothing more than a plan. This is evident in the way it's described, using a nonfinite complement clause: *John started to run a race*; *John decided to have a meal*. The participants in the scenario represent the scenario as a *goal* of theirs, rather than as a fact. Somehow, the representation has to identify the scenario as a goal *of theirs*, rather than of someone else's.

[29] Note this is very much like what happens in my model of motor learning: at the consequent state of a grasp action, when the agent has a stable grasp of the cup, an association is made between the desired state and the actual state.

[30] That implies that there's a 'mirror system for plans'. Several people have proposed this, but there's not much hard evidence for it.

I think it's useful to be guided the linguistic expression that describes initiation of a plan. For instance, *John and Mary decided* to have a meal. To do this, John might have to ask Mary *Would you like to [have a meal]? / Do you want [to have a meal]? / Shall we [have a meal]? / Let's [have a meal]!* In all these constructions, there's an explicit nonfinite clausal complement. The main clause features the participants *as agents*, along with some modal verb like *like* or *want*. In our model of clausal complements, the main clause effects a situation update, so the situation in which the proposed scenario is represented is *specific to the participants who will take place*: one situation will represent *John and Mary's* participation in a dining scenario, and a different one will represent *Sue and Jim's* participation in a dining scenario.

Now say I ask myself, as an observer, *What are John and Mary doing?*. (Or maybe someone else asks me.) The question refers to the current moment in time. But at this moment, there may be several other agents in the middle of other activities. Somehow this question must focus my attention on *John and Mary's* plans, and I answer my question by focussing on *John and Mary's* actions. Here's what I think happens. (1) I activate *John and Mary* as the (collective) agent in the WM episode medium, and *do* as the action. (2) Then I do a situation update, and clear the WM episode medium. The new situation represents John and Mary's *current operational plan*, I think. (Not their desire—but the plan they are currently executing by their actions. I know they're doing something, and I want to know what this is.)

What I want to find out is the *scenario* associated with this new situation. This is a matter of *inference*, of course. It's a well known fact that observers of an agent's action can reason ('abductively') from their observable behaviour to the likely plan or intention that drives this behaviour. The observer must perform some inference about the scenario John and Mary have adopted, by watching their behaviour: that is, their first-order actions, represented as episodes. (Note it's *John and Mary's* behaviour that I watch, rather then other things going on in the world.)

I suggest the key mechanism supporting this inference is one where the active situation unit(s) functions *as a query to the scenarios medium*. To start with, the observer doesn't know what the active scenario is, and leaves this blank. (Or perhaps activates a probability distribution of possible scenarios.) But he can observe John and Mary's first-order actions. As successive actions are observed, the situation SOM representation progressively comes to resemble the representation associated with a particular scenario. Eventually, the situation SOM unit allows a specific scenario unit to be reconstructed with high confidence. And then the observer has the answer to the question *What are John and Mary doing?*.

This reconstruction process is a very nice model of the abductive inference that an observer can make of the 'plans'/'intentions' of an agent, based on this agent's actions.

The answer to a 'What-is-*A*-doing?' question An important part of the above story is that *scenarios are represented in a way that abstracts over participants*: the same restaurant-meal scenario unit is used to represent John's participation in a restaurant-meal and to represent my participation in a restaurant-meal. The unit that represents actions, but abstracts away from their participants, is reminiscent of structures in syntax. For instance:

- A clause with a PRO subject. E.g. *PRO_i* having a restaurant meal is great fun. (*/Pro_i playing chess with $yourself_i$ is challenging.*)
- Nonfinite clauses with pro subjects. E.g. *$John_i$ wants/decides [pro_i to have a restaurant meal].*

In fact, a nonfinite clause with a pro subject is involved in the analysis of a regular present or past sentence. This is explicit in Old English: *John does [eat at a tavern]. Do* in Old English is an auxiliary verb that takes a nonfinite complement. The auxiliary *does* is supposed to convey (present) tense, and the main verb *eat* conveys the action. In slightly later English, an episode viewed as ongoing (i.e. aspectually progressive) is realised in the simple present: *John eats at a tavern.* (In French, the simple present still expresses a progressive.) The LF analysis in these cases is the same: tense and the open-class action are contributed by different projections at LF, but pronounced together at PF, by an inflected verb.

The above analysis suggests that the 'tense' head conveys something about a *plan*, or *scenario*, rather than just about tense.[31] If I say *John did eat*, I'm asserting that the whole 'eat' action was *accomplished*: but I think as part of this assertion I'm also asserting that it was John's *plan* to eat: that he took on a certain scenario, and saw it through to completion. That is to say, I think both present (progressive) and past (completed) epiodes make reference to a scenario.

Selection of a scenario, versus selection of an episode Let's say John makes himself the agent, and now decides what to do next. If we ignore scenarios, his decision is implemented through a distribution of possible episodes in the c-ep medium. He locks in himself as agent, and then refines this distribution, picking a patient (if there is one) and then re-refines, picking an action. But in the above discussion about scenarios, there was a new idea: after an agent selects himself as agent, *he selects a scenario.* He doesn't just have to choose a first-order action: he can also choose an *enduring cognitive state* to adopt, from a set of alternative enduring cognitive states. In psychology/neuroscience, this amounts to *task-switching.*

In our model, we can envisage a set of *candidate scenarios* being active after at every point when the situation SOM is updated. There are two possibilities: one is that the agent has no currently active scenario, and is considering which scenarios (if any) he should adopt

[31] If we follow this idea, the term 'scenario' must be able to apply to things other than plans of agents: I can say *The stone doth roll into the forest*, or in modern English *The stone is rolling into the forest*, without implying any intentionality. However, this progressive still conveys expectations about how the current scenario will *proceed* 'if things go as they normally do'. So I think a scenario needs to be defined more generally: some scenarios can involve active agents pursuing plans, but others just involve predictable sequences of episodes. I think this is a good idea.

now. The other is that the agent is in the middle of some task, and *interrupts* it—or even abandons it—because something more important comes along. I'll discuss these in turn.

If the agent has no active scenario, and is considering what to do,[32] the existing scenarios compete for selection on a number of grounds. I think these mainly relate to the emotional associations of the scenario. From a linguistic perspective, clauses with a PRO subject can certainly function as the subjects of sentences predicating emotions: *[PRO having a cafe meal] would be lovely*, *It would be nice [pro to have a race]*.

If the agent is in the middle of a scenario, then of course the *current* scenario will be one of the possibilities. Maybe there's a constraint in place to ensure that switching to a scenario while you're in the middle of some other scenario is only done under special circumstances. The special interrupting operation needed to change the active scenario is the kind of operation envisaged for the 'reorienting network' of Corbetta and Shulman, so I'll call it **reorienting**.

Several things combine to induce reorienting. A key idea is that it's driven by external events. One way to model that is to use the device of a situation SOM unit being used *as a query to retrieve a scenario.* We saw how that happened in the above discussion about plan inference during action observation. But it could also operate while the agent is *in the middle of his own plan*, to identify that recently-perceived episodes are better represented as part of a *different* scenario. This happens in particular when an *unexpected episode* is experienced. If something *unexpected* happens in the middle of my restaurant meal, one option is that I don't record it *as part of my restaurant meal.* There might be some other scenario that accommodates it much better: *and a retrieval cue will identify this other scenaio.* One option is that I 'save the state' of my current scenario (by associating the current situation with 'now', as described above), and activate the scenario which is more strongly associated with this current situation (i.e. which is better retrieved by it). A second factor should be something to do with emotional connotations of situations. If something unexpected happens that doesn't elicit a strong emotional response from the agent, he is likely to ignore it. But if it has strong emotional associations, then he's likely to switch to a new scenario.

There are a couple of nice aspects of the above story. One is that it chimes well with what people write about task-switching in neuroscience: for instance ideas by Schultz *et al.* that dopamine responses are associated with *unexpected rewards*, and Braver and Cohen's idea that unexpected rewards jolt the agent out of his current cognitive set. Another is that it makes a nice suggestion about how an agent *starts to learn new scenarios.* Basically, we can suggest that a new scenario is chosen when an *unexpected situation* arises. The process for starting to learn a new scenario could actually be something like the process of activating a new LTM individual unit. Say an unexpected situation arises, and we use this as a retrieval cue into the scenarios SOM. If the scenario is similar enough to an *existing* scenario, then you can activate that. But if not, there can be a different routine, whereby

[32] I'm considering scenario selection from the perpsective of the *agent* here: i.e. it's a form of high-level action decision, rather than an inference to a good explanation for the observed actions of some external agent.

you *select an new, unused unit in the scenarios SOM*, and activate that instead.

Reference to scenarios in language Scenarios can be referred to by nonfinite clauses (*I want [to have a restaurant meal]*) or by continuous clauses with PRO/pro heads (*[Having a restaurant meal] is fun*). They can also be referred to by straight nominalisations—and often are—e.g *I want* <u>a restaurant meal</u>, <u>A race</u> *is fun*, *John organised* <u>a race</u>. It's through these references that we can express *episodes* relating to the participation of individuals in scenarios.

Of course, other episodes *do not* refer to scenarios. When I say *John grabbed a cup*, there's no scenario: the cup is a concrete object, and this sentence describes a simple concrete action, that might well be—in fact almost certainly is—a *component* of a larger scenario like drinking or making tea.

But even actions that are just *moderately* less concrete than grabbing a cup are best off modelled as scenarios. Consider *John ate a sandwich.* This is not an atomic action: there is much reaching, chewing, swallowing. And the whole thing could be interrupted by all manner of other things: conversation, looking around, wandering about. Moreover, the object of the 'eat' action is emphatically not a simple concrete object. The object is in fact known to carry a reference to the action, in quite complex ways: *ate* is a count noun, rather than a mass noun, tells us that John's action of eating had a well-defined termination point, and that this point was reached (see e.g. Dowty, 1979; Krifka, 1992). In this case, the scenario is not referred to by the object noun, but by the verb *eat*.

This is a different way of referring to scenarios in language. In this case, the verb directly refers to a scenario, and the object noun specifies one of the 'participants' in the scenario. (I guess the sandwich here takes on a particular role, namely 'food'.) A simple implementation of the 'eat' scenario would say, 'iteratively take a bite of the food; chew it until it's mush, then swallow; when the food is all gone, you're done'.

To open the way for additional levels of hierarchy, each episode in a scenario can refer to the activation of another scenario.

8.11.6 The nature of participants in scenarios

All scenarios involve participants. These can be inanimate objects, behaving in temporally extended but predictable ways (as in *The stone was rolling into the trees*), but more typically, they're sentient individuals. In the typical cases, a scenario is a plan that an agent takes on, that involves a persistent cognitive set. The plan can involve multiple agents, in which case they have to have methods for agreeing on the plan, and allocating themselves roles. In this section, I'll discuss in general terms what kinds of participant roles we must envisage.

I have a nice model of participants in simple episodes: there's a (proto-)agent and a (proto-)patient, and these are defined in relation to a sequentially structured sequence of

SM operations. There are attentional operations that establish first the agent, and then the patient: and these are preconditions for executing an action. How much of this idea carries over to the participants of scenarios?

For scenarios involving sentient participants, I certainly think that the participants must be established at the outset, and that this strictly precedes any episodes happening 'within' the scenario.

Is it possible to envisage a natural *sequence* of participants, with these being attended to one by one? Now I'm talking not about the sub-episodes that take place *within* the scenario, but the meta-level/preparatory actions that *establish* the scenario's participants.

- In some cases, yes: where the scenario just involves one sentient agent doing a plan, the agent is definitely the first participant.

8.11.7 Reference to scenarios in WM episodes

In the above proposal, an important idea is that a WM episode can describe the process by which a given scenario is initiated—including identifying the participants of the scenario, so these can be bound to individuals. I think WM episodes can refer to scenarios in several different ways: in this section I'll enumerate some of these.

Reference to scenarios in the ACTION field For instance: *John ate a sandwich.* This sentence describes a (completed) scenario, not a simple episode, as just noted above. In this case, it's the ACTION field of the WM episode that identifies the scenario. I suggest the scenario has agent and patient roles, just like a WM episode. The AGENT field of the WM episode indicates the agent who is executing the scenario.

The PATIENT field indicates some *other* participant in the scenario. I'll call this the **external participant**. The scenario may change this participant in lots of ways: it must be defined in a way that survives all of these changes—and that in some cases, actually *drives* the changes. I suggest the agent always implements some *attentional process* associated with the patient. This might involve different things. In one case, it might involve keeping a record of the patient in the WM individuals system: for instance, the diner at a restaurant should register the waiter as a WM individual, and the waiter should register the diner.[33] In another case, it could involve keeping a *spatial pointer* on the patient, that persists over changes to it in various different ways. For instance, someone eating a sandwich can put a pointer on the whole sandwich, and keep it on the sandwich as it diminishes: the end of the scenario is when the sandwich is gone. Or again, an agent chasing Mary can put a pointer on Mary, and *track* her as she moves location. Or again, an agent mowing a lawn can put a pointer on *a place within* the lawn, and systematically 'explore' the whole lawn by moving this pointer, until it has visited all parts of the lawn. In this latter case, it's perhaps not appropriate to talk about the scenario involving a discrete *sequence* of episodes. There's one 'mowing' action, which is causing the mower to

[33] If a scenario is interrupted, perhaps these changes should be stored in LTM as well.

move: the structuring role of the scenario is *purely* in guiding this locomotion action by an attentional pointer.

The case of creation scenarios is worth discussing separately; I'll do that below.

Scenarios involving iterated actions In *John ate cake* or *John squashed bugs*, we have a scenario involving a single *iterated* action. When the patient of is expressed as a mass noun, I suggest the agent's attentional routine is different: the agent establishes a group or mass stimulus that has no global form, and then acts on the most salient portion of this stimulus (which could be a piece of cake, or an individual bug, or a group of bugs—whatever's easier), *without IOR*, thus without systematic exploration. Each iterated action is done on the portion of the stimulus that most readily *affords* the action: the most squashable bug, the nearest bit of cake. (In some cases, for instance drinking beer, or squashing bugs in a line, there is a 'flow' within the stimulus, which sets up an ordering on portions: but if the flow has no end, there's still a natural interpretation for an action that has no endpoint.)

Creation scenarios For instance: *John* <u>*made a cake*</u>. For 'make' actions, the patient identifies the thing that will be made: for instance, a cake. This process has already been described: see Section 8.9 for a summary. Creation actions are implemented in scenarios:[34] static assemblies that impose a sequentically structured action routine (Section 8.6), aided by hierarchically structured 'goal individuals' (Section 8.6.5).

The ideas I've just been discussing about how participants of scenarios are represented fit very naturally with this idea. Firstly, the PATIENT identifies *which scenario is to be activated*: there are different scenarios for making different types of object, and these bias the agent towards different sequences of actions. (The 'make' action in the ACTION field serves as an *indicator* that the scenario is identified by the patient, as is discussed more below.) Secondly, this model gives an account of the attentional processes associated with PATIENT. At first sight, there can't be any such processes for an action of creation: the object to be created doesn't exist, so how can it be attended to? However, in my account, actions of creation are guided by *goal individuals*. A goal individual is active from the very start, and can have hierarchical internal structure: the attentional processes are initially those associated with the goal individual. Progressively, as the creation act proceeds, an attentional processes develops on the thing being created, in parallel with the process on the goal individual, so that the created individual's properties can be systematically altered to conform to those of the goal individual.

The only special thing we need to add about creation scenarios is that the PATIENT field identifies a scenario *and also a goal individual*. Goal individuals are activated in a separate medium, so we have to add this.

(At the same time, you can also 'make' *episodes*. This is explicit in languages like French, where *make* is the light verb that has the role of the English light verb *do*. Is there some analogue of goal individuals for ordinary scenarios?)

[34] They were called 'LTM situations' in that earlier section: I need to standardise terminology.

Causative actions The causative actions discussed in Section 8.6.3 must sometimes be modelled as scenarios rather than simple actions. To open a door (i.e. to cause the door to be open) might be possible in a single simple action, but to clean a car (to cause a car to be clean) requires a structured programme of actions: one needs to systematically clean every *part* of the car. The attentional routine is the same as the one for mowing a lawn: exhaustive attention to every part of the car, accompanied by the same action. The only difference is that it's a causative action.

Reference to scenarios in the PATIENT field For instance: *John and Mary had a meal.* In this case, the PATIENT field of the WM episode identifies the scenario. The content of the ACTION field *indicates* that the patient plays this role. It's occupied by a special closed-class verb—a light verb.[35] Examples are *have [a meal]*, *do [a dance]*, *give [a presentation]*, *make [a cup of tea]*, *take [a holiday]*, *play [a tune]*.

The AGENT field of the WM episode obviously indicates one (or more) of the participants of the scenario. One possibility is that it always indicates the participant *whose plan will be represented* by the scenario. When it comes to thinking generally about scenario roles, and rules for linking these to episode roles, an interesting observation is that *different scenario participants sometimes need different scenarios*. The scenario implemented by a waiter in a restaurant has to be different from that implemented by a diner. A diner couldn't automatically fulfil the role of a waiter. Neither could a student automatically fulfil the role of a teacher: these skills have to be learned separately. Of course, a diner still needs to represent (some of) the waiter's *actions* in his scenario, because his own actions need to be coordinated with those of the waiter. But that's very different from knowing how to *be* a waiter. On this model, the referring expression 'meal' actually refers to a set of (one or more) scenarios, one for each participant type—and the verb indicates which of these versions is to be selected.

There can still be *other* participants in the scenario, as well as the agent whose scenario it is. For instance, the diner is a third-party participant in a waiter's 'meal' scenario; while the waiter is a third-party participant in a diner's 'meal' scenario. These are schema-specific, as described above.

Control verbs For instance: *I told John to eat a sandwich*, *I persuaded John to eat a sandwich*. The important thing about these is that they're sentences that describe the process whereby multi-agent scenarios are initiated. They're never thought of this way by linguists, of course, but I think it's a helpful angle. The former sentence describes an action that occurs during the process of establishing a scenario; the latter describes the result of a successful scenario-establishing action.) Here's another example: *I promised John to eat a sandwich on Tuesday*. This is another scenario-establishing action.

[35] I need to go away and read up Kratzer (1996) on voiceP, and Pylkkänen (2008) on light verbs and applicatives. My 'cognitive' model fits really well with those syntactic models.

One issue here regards the 'exceptional case-marking' that gives object case to the NP that provides the pro-referent of the embedded clause (*John* in the above examples). This suggests that *John* is part of the matrix WM episode. Indeed the physical actions of telling, persuading and promising involve a real direct object. But there must be verb-specific rules about which participant is understood as the subject of the embedded clause.[36]

8.11.8 The patient of a scenario: some ideas about patient properties and consequent states

Think about *John eats a sandwich.* Over time, the sandwich changes its *properties.* In what way? The easiest way to describe these changes is in terms of the effects of the eating action: in intermediate stages, the sandwich is 'partly eaten', and at the end it's 'eaten'. The same is true with *John mows the lawn.* Halfway through mowing, the lawn is a partly-mown lawn. At the end of mowing, it's *mown.* A lever can be *bent* (a lot or a little); a glass can be *broken.*

These 'intermediate properties' of patient participants are most easily *described* with reference to a particular action, but they are *perceivable* in their own right, even if you don't witness the action that brings them about. This is because of your *experience* of sandwich-eating/lawn-mowing etc. The properties of a patient object at intermediate states during execution of a scenario are crucially important for the execution of scenario: the agent needs to *monitor* them, so the scenario will certainly make some reference to them. ('When the patient object is like this, you should do this', and so on.) (The monitoring connects to the 'attentional tracking processes' I was discussing above.) In other words, it's information compiled into the *scenario* that allows you to look at a half-eaten sandwich, with no agent in sight, or a half-mown lawn, and recognise from its properties what's been done to it. I think in fact the half-eaten sandwich may *activate* the 'eating' scenario. (This is in the context of a model of properties, of course. Probably we first recognise the sandwich as a sandwich, and then do property-level IOR, and find out that this sandwich differs from the prototypical one in that it has featured as the patient of an 'eat' scenario.)

The sub-stages of a scenario: intermediate situations and intermediate properties When I take a bite of a sandwich, this episode generates an update in the Situation SOM. The new situation is one which 'predicts another episode'—obviously, in this case, another bite. But there's also an important *stative* component to the new situation. Remember: *LTM situations need to be involved in axioms about the changes that actions bring about on objects.* The effect of the bite action on the properties of the sandwich is implemented in an association between the new situation and the bitten sandwich. I think 'the bitten sandwich' should be represented as an RPC (since these things have to be detectable purely through perception). In fact, I suggest there should be many different

[36] In some cases like 'want', the direct object still has accusative case, even though it can be within the scope of the modal operator (e.g. *John wants a unicorn to do the washing up*). I don't have a story for that: my best guess is that this is one of the things that's learned by the word-sequencing network.

RPCs associated with sandwiches of different degrees of bittenness. I think there should be RPCs associated with sandwiches which *afford* different subsequent bites. (One might have a tempting next bite to the right, another to the left. These are real! They influence how the sandwich is eaten!)

I think these associations between RPCs of partly-eaten sandwiches and situations are what allow the properties of a partly-eaten sandwich to be described with a past participle like 'eaten'. Situation units have associations with RPCs—but these run in both directions, so if the RPC becomes active through perception, without any action monitoring, it can still activate a situation unit—and through that, a scenario unit (e.g. 'eat')—and through that, an 'eat' WM episode. In this case, since the sandwich appears as the PATIENT in the WM episode, we get a past participle, I guess. (No coincidence that in lots of languages, the past participle and the passive are the same.)

A new perspective on scenario participants One hypothesis that's suggested by these ideas is that the 'patient' scenario participant is represented in a special way within the scenario: *all the intermediate situations in the scenario are associated with representations of this patient participant, that encode how the scenario changes it.*

Now we have some interesting generalisations about 'agent' scenario participants and 'patient' participants. The agent participant is the one *whose scenario it is*: the scenario is deictically referred to a prior action that takes the perspective (i.e. adopts the mental state) of the agent. The patient participant is the one whose change-of-state is documented by the sequence of situations that the scenario records.

Patients of object creation scenarios The patient of an object creation scenario (e.g. a cup of tea being made, or a lasagne being made) really can't be thought of as an object with progressively changing properties. There is no object to start with. I'm modelling this one by referring to the properties of a goal individual, which are inspected. I think these might have to be linked to sub-situations in a creation scenario too, in conjunction with intermediate-stage properties of ingredients/components of the created item. For instance, when making a lasagne, at intermediate stages, at specific points in the recipe, there are indications about what the objects are that you have created, and their properties. You need a thick, smooth, salty béchamel; you need a well-reduced ragù; you need softened sheets of lasagne; you need finely grated parmesan. The main difference here is that there are several separate things that need to be made, before they are brought together. This information is in the scenario, somehow. So there are several points in the scenario where reference is made to perceptual properties, and executing the scenario involves monitoring 'patient' objects for these properties. (For instance, 'Heat the sauce and stir *until it is smooth and thick*'.)

It's interesting to think of the predicative adjective phrase *made of*, which lists the ingredients or components of a constructed individual. It does it in a way that refers to the creation scenario, I think.

Other agent participants? Another class of participant are collaborators in joint actions: for instance, the waiter (for a diner), or the diner (for the waiter). I've established that we need separate scenarios for dining and waiting: but I'm not sure how the collaborator is represented.

8.11.9 Representing locomotion actions in the scenarios framework

This idea of progressively changing properties can be applied to locomotion actions. Here's a simple example:

(8.4) John walked [to the door].

Recap about locomotion actions Recall from Chapter 5[37] that a locomotion action involves actions in two motor systems: one centered on the agent, and the other centred on the agent's environment. The action of *walking* is defined within the agent's frame of reference: it's an iterated act, controlled by a central pattern generator, where the agent moves his legs/feet to particular agent-centred goal motor states (and moves his arms in counterphase). Assume the act is parameterised: there are different parameter settings, which cause the agent to move in a straight line, or along various curved or discontinuous trajectories. The specific trajectory the agent travels along is specified in a higher motor frame of reference, centred on the agent's environment: in this case, let's say a room with a door in one wall. The higher-level motor system acts by imposing particular parameters on the agent-centred walking act. The parameters can change, so that the agent describes a trajectory with different degrees of curvature at different times, or with discontinuities in it.

Reconstruing locomotion actions within the scenarios/situations framework A locomotion action is an example of an action that requires the establishment of an enduring cognitive set: it takes time (often a lot of time), and the agent has to resist lots of potential distractors along the way. It's definitely the kind of thing that should be represented with a scenario unit. There should be a scenario unit representing the 'go' action. (Or more likely, several such units.) The scenario unit should involve a 'sequence' of substeps, which are simple atomic actions—in this case, these atomic actions are *literally* steps. (The important thing is that these steps can be executed with different parameter settings, so that the agent can steer around.)

Each action should update the 'situation' SOM, to represent the progress that has been made so far. These updates should be informed by a large amount of prior learning, encapsulating the agent's expertise in this kind of action. So when the agent locomotes to the door, the action will be 'guided' by a sequence of active units in the situation

[37] I can't point to the right place, since the chapter is still in draft form. But this summary summarises what *should* be there…

SOM. Each active unit will generate predictions (actually *recommendations*) about the next atomic action. In this case, they will be recommendations about particular parameter values for the 'walk' action, rather than about alternative *types* of action. And the atomic action taken at one moment will inform the way the situation SOM is updated—i.e. which SOM unit is picked next in the sequence. (This influence comes alongside the SOM's own recurrent connections, which can suggest particular trajectories by themselves, even without 'feedback' from the episode that just occurred.)

As discussed above, each situation unit in the sequence of units that encode the progress of a complex action is linked to the WM medium that holds object properties: this allows situations to be associated with gradually changing properties, brought about by an action that's under way. For locomotion actions, the gradually-changing properties relate specifically to *location* (in the allocentric medium already described). Thought of this way, *the situation SOM can be incorporated in an interesting way into a model of the hippocampal cognitive map medium.*

Firstly, the cognitive map must support a representation of *trajectories*: standard *sequences of locations*. If situations in the situations SOM are associated with locations in a map, the SOM's recurrent connections encode *sequences* of locations in the map. This is nice, because the situation SOM definitely holds representations of learned actions. (In a scenario like mowing the lawn, individual situations hold states of partial-mownness, but sequences of situations also straightforwardly hold *trajectories on the lawn.*)

Secondly, *the account of scenarios can be linked to an account of LTM environments.* Recall from Chapter 2 that the topological/topographical structure of the map of place cells is established by the currently active 'LTM environment' unit. LTM environment units hold gestalt representations of the shape of particular environments: the currently active unit supports a particular set of afforded trajectories. An LTM environment is very much like a scenario: it's a particular, static influence on the medium holding map locations. I don't think LTM environments *are* scenarios: they're representations of place, rather than time, which is very important. For the same reason, I actually think that place cells should be modelled in a medium that is functionally analogous to, but *separate* from, the situations SOM. The map of place cells should be activated in a special way by *perceptual representations*, as well as by recurrent inputs and efferent copies of the agent's actions. The perceptual pathway comes from ambient vision, and focusses on representing barriers in the local environment, and obstacles to navigation within it. This pathway was discussed in Section 2.5.2. It implements a complex mapping from retinotopic to allocentric location representations, that has to be learned. There's no equivalent perceptual input feeding the situations SOM: this is activated in a very circumscribed way, by a representation of 'the WM episode that was just experienced'.

On this model of 'separate but parallel systems', when we start a locomotion action, the top-down mechanism controlling activity in the WM episode medium *stops* being the situation SOM informed by the scenarios medium, and *starts* being the cognitive map (holding situation-like units) informed by the LTM episodes medium. This is better from a conceptual point of view. It also allows a nice account of the denotation of spatial PPs when they feature as verb arguments: we can say a spatial PP functions as a signal that control has passed to this special, separate, location-based higher-level control system.

Thirdly, and perhaps most interestingly, I think the model of how scenarios and situations interact may be able to implement an existing model of how the function mapping retinotopic perceptual inputs onto allocentric locations is *learned*. This is the model of Wiskott and Sejnowski (2002). In their account, the crucial constraint on the system holding allocentric representations is that the representations within it should *change slowly* as the agent moves location or orientation, despite the fact that the agent-centred perceptual representations from which it is created vary wildly with these movements. This constraint is very analogous to the constraint that requires *the same scenario unit to be active* during each situation with a plan. Assume the locations/environments system works exactly the same way as the situations/scenarios system, *except for additional input from the perceptual system*, mapped through a function that has to be learned, as just noted. During learning, the agent does actions that move him a small distance in space, and/or that change his orientation. These movements bring about big changes to perceptual representations, but they also occasion new values for the location and LTM environment units, through an update mechanism structurally similar to that for situations/scenarios. If we require during learning that the 'next' location unit that is activated should as often as possible be the same as, or at least overlap with, its previous value, and that the 'next' environment unit should likewise tend not to change, I think this should train the mapping function to convert perceptual representations into allocentric ones, by the same principles as Wiskott's model. After optimisation, there will still be some sudden changes to both the locations units and the environments units. These correspond to points where the agent *transitions* between environments.

This last idea points up two other interesting links between the locations/environments system and the situations/scenarios system. First, note there is a good analogue of 'dead reckoning' within the situations/scenarios system. Information about the agent's allocentric location doesn't just arrive from the world; it's also updated by (efferent copies of) the agent's own actions. Recall that the next situation is a function of the previous situation *and also the current episode*; the current episode is the input that's analogous to 'dead reckoning'. Second, recall that in the situations/scenarios system, there is an 'extended attentional tracking mechanism' that is associated with one of the participants. There is a tracking mechanism of this kind in the locations/environments system: but here, what is tracked throughout the whole process is *the thing describing the trajectory*. In a regular locomotion action like *John walked to the door*, this is the agent doing the action. In a causative locomotion action like *John put the cup on the table*, the tracked object is the cup (moving onto the table).

Interactions between the situations and locations systems If the locations/environments system is structurally analogous to, but separate from, the situations/scenarios system, it's important that these two play nicely together. The situations system should be responsible for deciding *what* locomotion action should be executed, in any given situation. Different actions are possible at different times. (I presume the allocentric locations system will nonetheless propose actions 'bottom-up', if they are possible/afforded: not quite sure how that works.) Then the locations/environments system *executes* the relevant locomotion action. At the end of the action, *the situations/scenarios system updates to register the completion of the whole action.*

Another possibility is that the situations SOM updates *in parallel with* locations in the locomotion system, rather than discretely at the end. After all, when an agent locomotes, there may be things other than location that change: for instance, the agent can become tired. In fact, the agent can pursue independent thought processes while locomoting: while walking, he can be watching an episode, having memories, and all sorts of things like that: these are things that are represented within the situations SOM. So I'll say that the environment/location system can run in parallel with the scenarios/situations system. It only relies on this latter system to *initiate* locomotion actions, and to register their completion (as mentioned above).

A representation of 'here' The agent needs to keep a record of his current location—perhaps in a number of different environments. (Because he can be in a whole heap of nested environments.) I think there can be a unit representing 'here', that plays a role analogous to the 'now' unit for situations/scenarios, that points to several locations, each of which points to a specific environment. The 'here' location in the current environment is updated by the zeroing process mentioned above, so when the agent moves it's constantly refreshed, but the representations of 'here' are held in LTM, so they persist over long periods.

Parallels between spatial and plan/action-based representations It's well known that progress within a high-level action plan can be represented using a spatial metaphor. We talk about different *ways* of doing a task; we ask ourselves 'where we are' 'in' a task, or we say we're 'in the middle of' a task, or that we're 'getting through' a task; see Lakoff and Johnson (**?**) for a famous account.[38] The above ideas about using the scenarios system to represent allocentrically specified locomotion actions, locations and environments would also explain these deep-rooted metaphors.

8.11.10 The PP system and its relation to the location/environment and situation/scenario systems

As mentioned above, while PPs naturally describe the location/environment system, they can also describe the situation/scenario system. (We can talk about an agent being 'in the middle of' a task; we can talk about 'ways' to do a task, just as we can talk about ways

[38] We also use spatial metaphors when talking about ordinary actions: for instance, we can say John ate a sandwich, but we can also say he ate 'through' the sandwich, or even that he 'ate his way through' the sandwich. Actually, I'm not sure this plays on the same analogy: the action actually involves a real, physical trajectory. So maybe forget about this.

as spatial paths.) In this section I'll revisit the PP system, to think about what PPs point to.

8.11.10.1 PPs in relation to the location/environment system

A starting point is that spatial PPs describe SM operations that take place within the location/environment system. These operations can be a mixture of discrete operations that transition between environments (see e.g. Section 5.7[39]) and operations that monitor continuous trajectories within the currently established environment (see e.g. Section 5.1.3). For example, the PP *on the desk* describes a reconfiguration action, where we transition from the environment *containing* the desk (in which the desk is an object) to one of the surfaces of the desk as an environment. On the other hand, the PP *to the door* describes a continuous trajectory afforded by the currently active environment, which aims at a particular object within this environment (the door), and reaches this object; the PP *round the desk* describes a different type of continuous trajectory afforded by the currently active environment, which travels round a particular object within the environment (the desk).

As discussed in Chapter **??**, PPs contain both PlaceP and PathP projections. I argued in Chapter 7[40] that a PlaceP describes an operation monitoring a continuous trajectory in the currently active environment, and a PathP describes a reconfiguration action, updating discretely to a new spatial environment. In this section, I'll try and relate the earlier analysis to the new account of environments and locations, that parallels the account of scenarios and situations.

The denotations of PathP heads Somehow, a Path preposition like *to* or *round* or *along* or *across* or *through* has to denote a trajectory. It has to be represented in an allocentric coordinate system, because it can be applied to onesself but equally to an observed locomotor. My question is: what sorts of thing are these 'characteristic trajectories'? What's their general characterisation? (Because the it's the general characterisations that the different prepositions will denote.)

8.11.10.2 A recurrent SOM model of environment-centred locomotion

[Place-holder: this is where Martin's new model is described.]

A network that can represent multiple environments We need to allow the network to represent several *different environments* using the above principles. The key device for achieving this is obviously the environment medium. If we turn on a different unit in this medium—say $E2$ instead of $E1$—the network's dynamics have the potential to change completely. However, recall that the network's dynamics results from a *combination* of static inputs from the environments medium *and recurrent inputs in the locations/orientations*

[39] Actually there needs to be an earlier section, somewhere in Chapter **??**, describing purely attentional reconfiguration actions.

[40] Still not written!

SOMs. Let's say the agent recognises that he has just arrived in a new environment, with which he is entirely unfamiliar. I envisage he will find an as-yet unused environment unit (let's say $E2$) and activate that tonically: now he will explore the new environment, and build a dynamical representation of *that*. Of course, it's crucial that the learning that takes place in the recurrent connections of the two SOMs, and in the next-action prediction function, does not *interfere* with the dynamics that defines the first environment. At the same time, we expect there to be some principles of navigation that *generalise* over environments. It may be that a SOM is able to work out how to represent the general principles and the environment-specific principles, in a way that keeps these latter principles from interfering from one another.

How an environment representation can be learned: (2) The function mapping from visual representations of boundaries to allocentric representations

8.11.10.3 A SM interpretation of PPs

8.11.10.3.1 Overview I suggest that a PP describes a (rehearsed) sequence of SM operations, in the allocentric/locomotion medium, that *transition* from the currently active 'spatial representation' to a new 'spatial representation'. A 'spatial representation' is a tuple: (i) an active LTM environment, and (ii) active units in the locations SOM and orientations SOM. The transition can be purely attentional, or it can involve monitoring of a movement in space.

An **attentional transition** is reported in two kinds of sentence. One is a stative (predicative) sentence describing something's location: for instance *The dog was in the garden*; *There was a dog in the garden*. The PP here features as part of the predicate. I suggest it describes the transition between the observer's *currently active* spatial representation and a new one. The other is in sentential adjuncts like *In Switzerland, everyone obeys the law.* The PP here establishes a new spatial context that has *no* direct link to the current one. The high syntactic position of the PP means it doesn't pick up a reference to the current spatial representation and transition from that: the PP describes a transition to a spatial representation identified through a *memory cue*, rather than through a specified transition from the currently active spatial representation.

A **physical transition** describes an action whereby something moves from one spatial state to another. The something can be the agent himself, or an observed agent (or inanimate object). In a sentence describing a physical transition, the PP features as an argument of the verb. In sentences describing locomotion actions (e.g. *Ali walked to the door*), the verb's subject is the thing whose movement is described. In causative sentences (e.g. *Ali put the cup on the table*), the direct object is the thing that moves (actually, is *caused* to move).

I'll focus on physical transitions first—specifically, locomotion actions.

8.11.10.3.2 I assume when an object is attended to, its spatial representation becomes *active* in the spatial representation system. (The spatial representation system holds allocentric representations, rather than agent-centred ones.) In a locomotion action, the active spatial representation at the point when the spatial transition is executed is the one associated with the agent: for instance in *Ali walked to the door*, the spatial transition is executed at the point when attention is on the agent, Ali. Within the PP *to the door*, there is no reference to this agent. In syntactic terms, the reference is given by the syntactic context of the PP; in SM terms, the active spatial representation is given *deictically*.

The PP describes a sequence of three things, in three XPs. First, an action of attention within the spatial system. This can be to a specific location within the currently active environment (as in *to the door*), or to the currently active environment as a whole (as in *around the room* or *along the path*).[41] The head of this XP is a Case-assigning head. The specific location is activated as a *goal location* in the current environment; the environment is just activated as an environment. Second, a *control regime*, that steers the agent through the environment on a particular path in relation to the goal or selected environment. Third, the completion of this path: in this case, this occurs when the agent reaches the goal location in his current environment (the door).

blob

Notes A few things to make reference to here, within the chapter on spatial representations:

- The notion that 'boundary pairs' form core representational structures within the Environments system, defining 'sub-environments': see e.g. Section **??**.
- The idea that when you shift spatial attention to a new item (e.g. in the operation reported by an existential sentence) you establish both a 'subject' and its containing sub-environment: see e.g. Section 2.8.3.
- The idea that a PP doesn't report the 'uncategorised' salient stimulus in an existential sentence: or at least that the top-level XP in a PP has no linguistic realisation. What a stative PP reports explicitly is the sequence of environment transitions which can link the observer's *current* environment to a spatial representation of the salient object (that is to say, an allocentric representation). See e.g. Section 5.18.1.1.

8.12 Authoring and querying episodic LTM

[This section is out of date, and can be replaced with our CogSci2016 model of LTM.]

The above circuit serves to learn generalisations about how episodes follow one another, but it also serves as an interface to episodic LTM. When an episode is experienced, we create a WM episode. A key point is that this episode occurs at a particular token time: there's a new time unit that's different from any other ones.[42] It also occurs in a particular LTM situation. So the SOM unit that's picked in the current situation SOM can encode

[41] *John walked along the path* presupposes John is already *on* the path, I think.

[42] Exactly how this time unit is generated is another story: this needs to be specified in Section 8.4.

a particular time and situation as well as the content fields of the episode. Effectively, the current situation SOM stores episodes in LTM, as well as encoding a WM representation of the current context. Provided it has enough capacity, it allows a WM episode to be used as a cue to LTM, to retrieve specific contexts in which this episode was experienced. The mechanism is as follows. The WM episode activates a set of units in the current situation SOM. Each of these encodes either a situation type, or a situation token.

A unit in the current situation SOM that encodes a token situation can be used to retrieve a token LTM situation, context and time in which the episode was experienced, by downwards reconstruction. (Note that it retrieves a *tuple* LTM situation, context and time: and that's important, since these things relate to one another.)

A unit in the current situation SOM that encodes a generalisation over situations—i.e. a situation type—can also be used for downwards reconstruction. In this case, what we get back is a set of separate distributions in the context, LTM situation, and time media. Here we may have lost some dependencies between these distributions. However, LTM situations are also generic things, and there can be generic times, and contexts can be generic too—so we might be able to retrieve the generic situation in which a WM episode occurs.

8.12.0.4 ***When*** clauses

As famously noted by Moens and Steedman (1987), *When E1, E2* can signal various different relations between episodes $E1$ and $E2$.

First, $E2$ can be the episode that follows $E1$, as in:

(8.5) When they built the Forth bridge, traffic flowed much better.

Second, $E2$ can be an episode that happens *within* $E1$, as in:

(8.6) When they built the Forth bridge, they used the best materials.

Finally, $E2$ can be an episode that happens in the context that $E1$ occurred in:

(8.7) When they built the Forth bridge, their traffic problems were very serious.

Our episodic LTM circuit needs to be able to allow these same three possibilities. In each case, we present $E1$ as a query episode.

To get to the episode that follows $E1$, we find the best-matching unit in the current situation SOM, and activate it by itself, then update to get to the next situation, and find the most active unit in the candidate episodes SOM.

To get to an episode within $E1$, we find the best-matching unit in the candidate episodes SOM and activate the associated LTM situation unit, through the 1:1 connections between candidate episodes and LTM situations. We then use this unit to activate a unit in the current situation SOM—which will represent a context *within* the situation associated with $E1$.

To get to an episode that happens in the same context as $E1$, we find the best-matching unit in the current situation SOM, and by downwards reconstruction, activate the LTM situation/time/context that $E1$ occurred in. Then we see what the most active episode is in the candidate episodes SOM. (Alternatively, we take the time and situation by themselves, and see what *states* obtained in this time/situation.)

In each case, reconstruction of a particular $E2$ unit might also depend on a particular question. Questions often presuppose a particular context: for instance *When they built the Forth bridge, what materials did they use?* To process this question, we first establish an LTM context and then pose the question.

8.12.0.5 The operation of 'being reminded of a situation'

When E1, E2 can also report a scenario in which the agent is simply reminded of an episode $E2$, because of its similarity or relevance to something happening now. The LTM circuit has to support this reminding operation too.

The mechanism for this has to be a bit different from the one I envisaged before. Previously, I had the WM situation distinct from the LTM situation. But now they're in the same medium (the 'current situation'[43]).

I now have to propose that when a new current situation becomes active in the situations SOM, one possibility is that there are *other* units in the situations SOM, representing *other* situations, that are similar enough to the current situation that they compete to remove the system from experience mode, and put it into some form of LTM retrieval mode. This similarity-based competition is pitted against the new things arriving in the WM episode medium, I presume. If these are exciting enough, then we stay with the current episode. (Note that the candidate episodes SOM can happily represent a distribution over expected *current* episodes, even while the similarity-based competition happens in the situations SOM.)

However the reminding mechanism works, once we have reactivated a unit in the situations some representing some situation other than the current one,

8.12.0.6 A general account of *when*?

All these operations signalled by *when* have to have something in common. The basic commonality is that *when* signals a scenario where a new current situation is selected discontiguously, rather than by the regular recurrent update mechanism. But we can say more: the new current situation is always selected as a function of the WM episode denoted by the subordinate clause, without the help of the 'current' current situation.

I think *when* could perhaps be taken to denote the instruction 'clear the current situation medium'. Following that operation, the new current situation medium is a function *exclusively* of the WM episode. (Either directly, or via the candidate episodes SOM and then the LTM situations medium.)

[43] I think I should call it the **situations SOM**. And I think the LTM situations medium should be called the **structuring situation**

8.13 Relationships between episodes and times

8.13.1 'Now'

I suggest the time network has a special method for representing episodes which are occurring 'now'. Whereas normal times are constantly updated, 'now' is special, in that it remains constantly active.

I suggest that when an observer begins monitoring an episode, he links the token episode representation to the 'now' environment.

I suggest there are several 'now' environments, operating at different temporal scales. For instance, if I'm making a cup of tea, then I can add this whole episode to an environment representing 'the current half-hour' (or thereabouts). Within the process of making tea, there is a specific episode under way: let's say the kettle is boiling. So I can add this kettle-boiling episode to an environment representing 'the current two minutes' (or thereabouts). At this finer granularity, I can represent that some sub-episodes within the tea-making episode are in the past: for instance the episode of filling the kettle has already occurred.

8.13.2 Simultaneous episodes/situations

People can represent episodes (or situations) happening simultaneously. In language, the clearest way of signalling overlapping episodes is using the subordinators *while* and *when*, as in the following examples:

(8.8) While the kettle boils/is boiling, put tea in the pot.

(8.9) When the kettle is boiling, put tea in the pot.

(8.10) When the kettle has started to boil, put tea in the pot.

The subordinate clause in these cases is aspectually stative: it refers to an activity through which an episode will eventually be accomplished. My basic assumption is that the WM episode buffer can only hold one episode at a time: so such examples have to be dealt with on the premise that the observer *begins* perceiving one episode and then interuppts their perceptual process, saving a stative representation of the progressive activity, and then experiences a separate episode (or state).

8.13.3 Returning to an interrupted episode

It's important that the agent can re-establish an episode he was monitoring previously, and continue to monitor it. If the ongoing episode was interrupted and stored as a stative activity associated with a certain time period that extended into the future, then activating 'now' in stative LTM could yield a distribution over units linking objects to properties.

n he was in previously. I'm thinking that this can be done by linking LTM situations to LTM time units. E.g. you could link 'wider-situation' to a representation of 'now' at a certain granularity (e.g. ‘this-current-half-hour’), and then simply reactivate ‘this-current-half-hour’ in order to recall the situations that are ongoing at this time. (Note there could be several of these: for instance you could be unloading the dishwasher at the same time as making the tea.) Note also that these references to ‘now’ could also be of different grain-sizes: for instance at a finer temporal granularity, what I’m doing ‘this minute’ might relate to a finer-grained task: e.g. what I’m doing within the task of making tea.

The ‘now’ environment is the agent’s method for working out what current situations he’s in the middle of.

8.14 Modelling the hippocampus and consolidation

My current idea is simple: there are two context SOMs for representing episodes in LTM. One is implemented by the hippocampus: this one is trained immediately when an episode is experienced, and holds a detailed representation of token times, even for unexceptional episodes. The other is implemented ‘in cortex’: this one is trained offline, to avoid the problem of catastrophic forgetting.

Part II

Other sentence types

Chapter 9

Cognitive representations of causal relations

In this chapter I'll introduce an account of how causal relations are represented in the brain. I started writing it after finding the need for representations of causality in the account of feelings, see especially discussion of *The chocolate enraptured John* (Sections 11.3.2 and 11.3.3). But I'm now presenting it as a preliminary to that discussion.

9.1 Pure statements of causation: SM operations occurring within the cognitive inference mechanism

I don't really want to analyse *The chocolate caused pleasure to arise* as a volitional action: that feels completely wrong. It sounds to me like a pure statement of causation, something more like the following sentences:

(9.1) The heat caused blisters to appear in the paint.

(9.2) The heat caused blisters in the paint.

(9.3) The heat caused the paint to melt.

(9.4) The heat melted the paint.

In Example 9.1, what's caused is an episode: 'blisters appear [in the paint]'. In Example 9.2 what's caused is, on the surface, an object ('blisters in the paint')—but it's pretty clear that what's really caused is the *existence* of these blisters. So they're not that different. Examples 9.3 and 9.4 show that the episode that's caused doesn't have to be one in which an object appears: we can have normal causative actions being done by agencies that are nonvolitional. (As my reviewer pointed out.)

One possibility is that humans have a special cognitive faculty for identifying causally related episodes, something separate from any faculties considered so far.[12] I now want to think about the nature of this faculty.

The faculty is definitely one that learns associations between representations. And these associations have to have certain properties: heaps has been written about this. One of the key properties is that associations have directionality: there's a cause episode and an effect episode. (I'm presuming that the things related are episodes.) However, there are *inference mechanisms* that allow you to go in *both* directions. You can reason using *deduction* from causes to effects, and using *abduction* from effects to their probable causes. So: I think the general cognitive mechanism I'm now focussing on is an *inference mechanism*. But the thing I'm thinking about is a specific part of this mechanism: the part that identifies causally related episodes. Once you have identified these, then you can use various patterns of inference.

I can think of three contexts in which we can identify a relation of causation between episodes. One is where the episodes occur externally in the world. (E.g. the episode of heat arriving causes the episode of the paint blistering or the paint melting.) One is where the episodes relate to an agent's motor actions. (E.g. the episode of my performing a certain causative motor action causes the episode of the cup breaking.) One is where the episodes occur in my head, and are apprehended in feeling perception mode. (E.g. the episode of the chocolate stimulus arising causes the episode of my feeling of delight welling-up.)

I think the idea I like is that the word *cause* signals the establishment of a new mode whose function is to identify causally related things. Specifically (as briefly mentioned in Section 11.3.3), I suggest that *cause* describes the execution of a mode-changing operation that 'engages a network encoding cause-effect relationships'. I'm envisaging a special network in the brain *dedicated* to doing this, that has application in many cognitive mechanisms, and is designed to work alongside other mechanisms. (So, for instance, it's also engaged in identifying causally related episodes in the external world, and in executing causative actions.) What I want to think about is what this network looks like, and what operations are involved in turning it on. (Are these nonstandard SM operations?)

I'll first consider *The heat caused the paint to blister*. What's the SM process described by this sentence? What's the first thing that happens? One possibility is that the observer first apprehends the episode involving the heat, and then apprehends the episode in which the paint blisters, and then (somehow) apprehends that these two things are causally connected. This doesn't feel right to me. I think the time at which the causal connection between the episodes is identified may not be the time at which the episodes are actually experienced.

[1] Maybe this same faculty could be at work in normal volitional causative actions like *John broke the cup*. Which would make my reviewers happy. *John caused the cup to break* does have the same syntactic structure as *The heat caused blisters to appear*. Maybe volitional cause sentences recruit the cause-identification network *in the context of a motor operation*. I like this idea a lot.

[2] In fact the same faculty presumably underlies good old causative sentence/clause *connectives* like *because*, *since* and so on. So I know something about the nature of the faculty already. Even *If X then Y* might encode something originating from this faculty.

An alternative possibility is that the relation of causation always links 'second-order' representations (in Damasio's terminology)—in other words, that the relation is identified in a cognitive mode where semantic representations are objects in their own right, and the 'episodes perceived' are episodes in which semantic representations *activate.* To be specific: imagine there's a network whose input is an eventuality representation E_2 in situation S, and whose output is the eventuality representation E_1 that is likely to have caused E_2. There are many constraints on this. Firstly, E_1 must plausibly *obtain* in situation S, or right before it. (If it's not the case, or if it's itself unlikely, then it's a bad candidate to be the cause of E_2.) Secondly, there must be some sort of asymmetric *learned association* between E_1 (in S?!) and E_2.

This makes me think of another area of the model: the mechanism that implements *updates to the WM situation* (i.e. to the distribution of possible episodes in the candidate WM episodes buffer, and similar distributions in other WM media representing individuals, environments, etc). When you update from situation S_1 to situation S_2, the new things in S_2 can certainly be considered to be caused by things obtaining in S_1. I'm suggesting there's a network that allows us to identify particular *component* eventualities in S_2, and work out which *component* eventualities in S_1 are responsible for their appearance in S_2. And this would be the causative network.

Note: in order to identify a relation of causality between cause eventuality E_1 and effect eventuality E_2, it's very important to think about *onsets* of these eventualities. This again makes me think that causality has to be ascribed within an internal perception mode: because, as mentioned above, in this mode, the *episodes* that occur relate to the *activation* of mental representations (which precisely identifies their *onset*). This also makes me think that the WM-situation-update function is likely to be involved. Some things ('fluents', in logicians' parlance) can change: often they don't change for a long time. The frame axiom applies. But then suddenly they change. All these fluents are managed by the WM-situation-update function. For present purposes, the point is: when they *do* change, you want to ask why.

Here's a specific idea. The **cause-identification network** *watches* the WM situation: i.e. the candidate-WM-episodes buffer. It's a network that's driven by *changes* to the current WM situation. These changes occur *individually.* New things *start* being true, other things *stop* being true. When a change is detected, the cause-identification network is trained to map this onto the *previous* WM situation—I think in its *entirety.* After training on a range of situations, the network will be able to take any given aspect of the WM situation, and make *predictions* about what the relevant circumstances were in the previous situation that led to ('caused', by definition) its onset.

9.1.1 The cause-identification network and unaccusatives

Think about *The heat caused the paint to melt.* Quite often this is used in a close-perspective narrative as a way of introducing the paint melting as an episode that updates the reference time. The episode could also just have been introduced as *The paint melted.* A standard analysis of this unaccusative clause (see e.g. Levin and Rappaport-Hovav,

1995) would say that there's an implicit structure:

(9.5) [Something] CAUSED [The paint BECOME melted]

A lot of this fits with the way I'm now construing the SM process. The fact that there has to be a *change of state* involving the paint fits with the idea that the cause-identification network takes *onsets of representations* as its input. The paint didn't used to be melted, but now it is: this is something that will trigger the cause-identification network.

9.1.2 Deployment of the cause-identification network: a perceptual state-change detection mechanism

Think again about the paint melting. The observer *perceives* the paint melting. I'm suggesting he has to represent this operation within the cause-identification network, rather than in the regular WM episode network. He can either represent it with *The heat melted the paint/The heat caused the paint to melt*, or with the unaccusative *The paint melted.* In each case, he's *not allowed to represent it as an ordinary WM episode.* This seems very important.

Do the above considerations mean that the cause-identification network is the 'SM mechanism' that actually *perceives the paint melting*? What would that mean? Surely in one sense I physically (visually) *attend* to the paint, and (visually) *experience* its change of state from solid to liquid. Here's an idea. There's a special perceptual mechanism that hooks *directly* into the WM-situation-update function: specifically, it directly delivers the kind of state-change 'episode' that provides input to the cause-identification network. This could very well relate to a visual mechanism like object files. We look at an object, which is continuous over space and time (e.g. some paint). Some perceptual mechanism (e.g. relating to categorisation of dynamic stimuli, patterns of motion, or something like that) becomes active: we now know that a change is happening in the object being attended to.

Here's an idea: maybe the perceptual mechanism that detects a change happening right now in the object being attended to (i.e. a change of state in this object) is something that can *trigger* activation of the cause-identification network. (Remember we have to think of activation of this network as a mode-changing operation.) Axiomatically, when the observer is tracking/fixating a certain object, and a pattern of motion is detected in this object, we can assume a change of state is occurring in it, and we can represent the onset of the new state in the new WM situation.

9.1.3 The cause-activation network and DP/verb-raising: some questions

In the case I'm discussing, the two verbs are *cause* and *melt*. It's a VP-shells structure of some kind. The subject of the upper verb (*cause*) is unknown; the subject of the lower verb (*melt*) is *the paint*. For some reason, (i) *melt* can raise to adjoin *cause* (and get pronounced at PF as plain *melt*); (ii) *the paint* can appear as the surface subject, even though it's 'semantically' not a subject.

My guess is that the reason *the paint* can appear as surface subject is that the action of attention that gets everything going is *an action of attention to the paint.* This, presumably, is done in external observation mode. (The observer's attention is drawn to the paint, an external object, therefore he enters this mode.) So until here, things are happening in the WM episodes network. At this point, the perceptual state-change-recognition system activates, and turns on the cause-identification network. My original thought was that this network takes the just-identified state onset (the paint being now melted) and infers a cause in the previous WM situation (the one in which the paint was not yet melted). But if it works that way round, it's odd that the cause doesn't appear *second* in the SM routine. The order of operations in the cause-identification network seems to be (i) identify effect; (ii) activate the network; and (iii) make a guess (based on abduction) about the cause. But, at least when the causative sentence includes explicit reference to the cause, the cause comes first, and the effect second.

9.1.4 The role of the cause-identification network in a causative action

In *John crumpled the paper*, the volitional agent John does an action that causes the paper to crumple. Question: what role does the cause-identification network have in the execution of a causative action like this? (Or for that matter, in its observation?)

My idea so far is that the cause-identification network links a 'newly-arising' eventuality e_r in the WM situation at time t_1 to an eventuality e_c obtaining in the immediately previous WM situation at time t_0, identifying e_c as the cause of e_r. (Along the lines posited by Waldmann *et al.*, 2008: the network focusses on causal relations linking *individual* eventualities, for reasons of cognitive economy.) The network can be *enabled* (i.e. 'cause-identification mode' can be *established*) directly by a perceptual mechanism that I'll call the **state-change-detection mechanism**: the mechanism that classifies a pattern of movement taking place in an object. Somehow, this network becomes active when the agent executes a causative motor action. How does it get activated? And what role does it play?

One possibility is that the perceptual mechanism that perceives an episode in the external world in Jeremy's model is not a general purpose episode-perception mechanism, but something more specialised, namely the state-change-detection mechanism just mentioned above. (This chimes with Levin and Rappaport's idea that unaccusatives always involve a change-of-state.) If that's the case, then the cause-identification network will become activated. And *then*, a potential cause will be sought for the observed change-of-state. If I can set things up so that one of the things that's observed in the WM situation just *prior* to the one in which the change-of-state occurs is *the agent's causative motor action*, then the cause-identification network is in a perfect position to *identify* the agent's own action as the cause of the change-of-state—which involves ruling out *other* candidate causes, which might take place in the world. If the agent's motor action *is* identified as the cause, we're not done yet: the agent can now learn a *direct* connection between the causative action and the observed change-of-state, *within his motor system.* After this learning has happened, we have an interesting situation: when a causative motor action is executed, the cause-identification network will still 'objectively infer' that the motor action as the likely cause of the change-of-state, even though there's another part of the brain (in the motor system) which knows that it *really is* the cause.[3]

In fact, something very similar can be envisaged for causative influences of stimuli on feelings. I suggest that the process of *getting a feeling* is a change-of-state process, *that the perceptual state-change-detection mechanism can pick up.* (This is already an interesting idea: this 'perceptual' mechanism has an 'eye' pointed to the outside world, but also an 'eye' *pointed inward*, as Damasio would have it.) If this is the case, then when I experience the dynamic process of a feeling 'welling up' in me, this change of state (in me) activates the cause-identification network, which considers candidate causes for this (internal) episode in the preceding WM situation. And in the scenario I'm thinking of, identifies a particular stimulus (e.g. 'the chocolate') as a plausible cause. (Because other candidates have been ruled out, or judged to be less likely.) Therefore the cause-identification network learns a causal link here. But at the same time, I think there's a *direct*, *associative* connection that's made between the stimulus and the (pre-attentional) feeling. (This is the piece of 'learning' that's accomplished in the SM routine described by the sentence *The chocolate delighted John.*) So *after* this associative learning has happened, the stimulus 'chocolate' *really does* cause the feeling delight, as a *learned* response. But again, the cause-identification network doesn't know this. So it's a self-fulfilling prophecy. But again, it's good that the cause-identification network continues to be objective about this causal process, because the world may change: chocolate may start to elicit a *different* feeling, and something needs to be able to recognise ths change. This is what behaviourists would call the process of *extinction.*

9.1.5 The cause-activation network and DP/verb-raising: another try

I think an account of goings-on in the cause-identification network will form part of the SM model underlying four types of sentence: volitional causatives (e.g. Example 9.6), pure causatives (e.g. Example 9.7), unaccusatives (e.g. Example 9.8) and stimulus-as-subject sentences (e.g. Example 9.9).

(9.6) John bent the lever.

(9.7) The heat melted the paint.

(9.8) The paint melted.

[3] Actually the cause-identification network still plays a useful role in calibration and plasticity: the associations learned in the motor system between the causative action and the change-of-state might stop being true, and something needs to be able to recognise that, and learn some different ones.

(9.9) The chocolate caused-delight.

In each case, at some point (i) a perceptual modality identifies a change-of-state (either in the external world or in an agent's internal milieu); (ii) this triggers activation of the cause-identification network; (iii) a cause is sought for the observed change-of-state; (iv) if one is found, it is activated individually, and (v) the change-of-state is lit up simultaneously, so that a link between them is created in the WM-situation-update mechanism. [If it's not found, then there's an empty subject, in some sense.] Now I want to think about the differences between the various cases.

9.1.5.1 *John bent the lever*

In *John bent the lever* (seen from first-person perspective), John first attnends to himself, then attends to the lever as a target object. Then he 'executes the CAUSE action': according to my model with Jeremy, this is a special action somewhat like a mode-changing operation (under volitional control), in that it engages a new network. This network has two properties: (i) it generates arbitrary motor actions, just like the normal motor control network; but (ii) it is trained by the consequences of those actions. In the sense that if the motor action results in perception of a change-of-state in the lever—in other words if the action results in a meaningful update being registered in the WM situation—then an association is learned between the motor action and the perceived episode/update. In my model with Jeremy, both of these things happened in the same network. But now I think they should be seen as happening in separate, interacting networks.

A key point in the above suggestion is that within the WM episodes system, CAUSE is *construed as a motor programme*, even though it's a special mode-changing operation. It's like ENTER-VERBAL-MODE in this sense. And just like ENTER-VERBAL-MODE, as a *consequence* of activating the cause-identification network, the SM signals that are evoked *next* in the WM episode medium have a special characteristic structure. The first signal to occur is a change-of-state operation: in our example, *bend*. The whole point of the causative action is that it reliably brings about this change-of-state, so that definitely occurs next. (The change-of-state is reported as a SM operation, which is fine, since these are update operations.) There are two further questions.

Firstly: what interpretation can be given to the reafferent sensory representation that's activated as a side-effect of 'monitoring' the lever's change of state? I think that's fairly straightforward: the reafferent representation is a dynamic representation of the target object (the lever)—i.e. a representation of the lever as a particular pattern of movement. (So if that pattern of movement is seen again, the object representation 'lever' can be triggered.) On this definition, the V head denotes bending 'as an operation in the motor system', that transitions from one SM context to another. The Spec of V denotes bending as a pattern of movement characteristic of an object.

I now need to discuss how learning of causative actions happens. The main idea is, as for verbal mode, that the agent *experiments* with the operation of enabling cause-identification mode, just as he would with any other action. To begin with, the motor action conducted on the object in this mode is completely random. (Or perhaps it's random *modifications* of existing ordinary transitive motor programmes.) Every so often, the motor action *happens to produce a change of state in the controlled object.* (By this, I mean an external episode that's detected by the change-of-state-detection mechanism.) Because cause-identification mode is enabled when this change-of-state is registered, there's an interesting coincidence, which enables learning: somehow the change-of-state is by some axiomatic fiat *set* to be the thing that the motor action produced. Ideally, this would happen in the same way that in a grasp action, when there's a certain tactile stimulus, the current motor state is axiomatically set to be the goal state associated with the visual representation of the stimulus produced earlier. In the case of a grasp action, a certain tactile stimulus axiomatically indicates a learning opportunity: this is the time to learn a function mapping a visual representation of the target onto an affordance-based representation (i.e. a motor goal). In the case of a bend action, I think the change-of-state detected in the target object might similarly indicate a learning opportunity. In this case, what's learned is a function mapping *something* onto *something*: perhaps mapping the executed motor programme (i.e. the physical movement) onto the caused change-of-state.

But I think expressing the function like this doesn't quite capture what's being learned. The way causative motor actions are construed above means that we connect motor actions to *aspects of the current WM situation.* That's the interesting effect of the mode-changing operation effected by entering cause-identification mode. And if we think more broadly about planning, the aim in a mature planning system is to achieve a certain goal situation. The effects of a grasp action aren't represented with any reference to the objective 'situation'. They are defined in terms of *objects.* (The effect of a grasp action is that an object is in a certain location in relation to the observed agent's body, and that a certain tactile representation of the object is achieved. These are not fully *external* achievements: they relate to the observed agent's body. The effect of putting a cup on the table is that the cup does something that it could do even without the agent's intervention: it's a change-of-state *in the world* (i.e. undergone by something other than the self), that leads to a state *of the world* (i.e. unrelated to the self, defined in a coordinate system *separate from* the self), defined in the same representational medium that represents 'the current situation', in all its complexity. Situations, or rather *aspects* of situations, are now things that can become *goals.*

What does it mean to say that the consequent state of a particular objective episode is a goal? I think this can be defined in reference to the WM-situation-update function. Say a WM situation is a distribution of possible episodes, somehow each represented as the consequent state it results in. Assume we're in action execution mode: so the activation level of each episode in the distribution says something about the value of its consequent state to the agent. Some kind of very high-level learning decides on the vlaue of each episode in the distribution. During learning, the agent is in situation S_0 and does an action that results in a new situation S_1 featuring a newly-appearing consequent state cs_1. Say in this new situation the agent receives a reward. The newly-arriving reward can be associated with the newly appearing consequent state, and the agent can learn a piece of the WM-situation-update function. Namely: if you're in S_o, World states are now things that can be goals, that can have value, and so on. And t this point, when an agent experiences a reward in a given situation,

9.1.5.2 *The heat melted the paint*

One possibility is that the operation activating the cause-identification network causes the contents of the WM episode buffer to be *cleared*, so that the sequence of SM operations represented in this buffer represent in a pure way what goes on in cause-identification network. (This emphasises the ability of mode-changing operations to reset the WM episode buffer.)

Another possibility is that the whole SM sequence is accumulated in a specialised WM medium for holding operations in the cause-identifying network, and read back from this medium. I think I prefer this idea, because it provides a more useful structure for a model of sentence planning. In this case, the observer would have the *option* of reading out the whole sentence from this medium, but might also have another option of reading out a different sequence stored in the WM episode buffer (see paragraph below).

9.1.5.3 *The paint melted*

Here the idea would be that we start off in external perception mode, attend to the paint, then represent its change-of-state as the verb *melt*, which activates the cause-identification network—however, in this sentence, the activation of this network is implicit, and the sequence of SM operations executed within the cause-identification-network medium is simply unreported.

9.1.5.4 *The chocolate caused-delight [in John]*

In here there are two novel ideas that need to combine. One is the idea that an undifferentiated strong emotional signal in the pre-attentional map of feelings causes establishment of **feeling-perception mode**. The other is that a change-of-state in the agent due to an emergence of *one particular feeling* (delight) triggers establishment of **cause-identification mode**. The structure of the sentence reflects these two operations, and how they interact.

A key thing here is that *early in learning*, cause-identification mode is only activated at the point when the feeling emerges. The mode then searches 'retrospectively' for a candidate cause. (The identification of a 'winning' stimulus is probably the outcome of this search.) As learning progresses, the cause-identification network discovers that the cause of the feeling is a certain active stimulus, and *associates* the stimulus with the feeling, so that the stimulus *axiomatically* generates the feeling. This has somewhat the flavour of motor learning: perhaps whenever you activate a stimulus in feeling-perception mode you also activate cause-identification mode

Here we first attend to John as an environment: so now we're monitoring episode involving activation of representations within John's brain. (That's the 'in John' part.) We're now in feeling-perception mode. We first identify a single stimulus

9.2 Aside: learning sequences of actions and macro-actions

[This section feels out of place! Where does it fit?]

Often it's important to do several actions, or to do a structure of actions, in order to achieve some distal goal state. E.g. Shallice and Cooper's classic example of making coffee: the agent has to [put water into the kettle, boil the kettle], [get a cup, put coffee into the cup], [pour water into the cup]. We might imagine a WM medium that stores sequences like this, similar in implementation to the ones encoded by syntactic structures, but reflected by discourse structure rather than sentence structure. This medium would clearly relate to the WM-situation-denoting medium, that represents updates in units of whole episodes.

Take a really simple example: we get a nice feeling when we grab a cup and then bring the cup to our mouth and then tip. So the feeling is associated with that state. Say we were recording the operations we did, one by one, in the WM-situation-denoting medium, the candidate WM-episodes (where I am assuming they're represented as reasonably localist units). We need a device that creates **macro-actions**: for instance, drinking from a cup involves raising it to your mouth and then tipping and then swallowing. Once we have macro-actions, we can reproduce our complex action strategies and encodings.

How do we get from [put water into kettle, switch that-kettle on] to [boil kettle]? I think the operation of *indexing* is going to be key here. A very similar operation to the indexing that links elements of plans in syntax-reflecting SM routines, e.g. that links WM individuals (prepared SM sequences in WM) to reafferent states of SM operations in an episode-denoting SM routine. There are a few points to think about here.

First, we're talking about SM routines stored in *long-term memory* (defined by synaptic weights) rather than working memory. Macros like boiling the kettle and drinking from a cup clearly endure over the timescale of LTM, don't have to be rehearsed to be maintained, etc. But they're still prepared routines. And many people have noted that WM and LTM can encode similar structures: for instance think about Burgess and Hitch (2005). If you hold/rehearse a sequence of operations in WM enough times, you start to create a *LTM* representation of the sequence. Viz telephone numbers; viz the Hebb repetition effect. So: I think macro actions *are* held in a WM medium: but by dint of repetition, they get encoded in LTM as well.

Second, we're talking about a collection of SM routines that all exist within *the same neural medium*. The indexing of a WM individual by a WM episode is a case where the indexing connects a SM routine in one medium to one in another. But I'm assuming all the stored SM routines that define action macros occupy the same (LTM) neural medium. This allows macros to get arbitrarily hierarchically structured: which is definitely what we want.

Third, it's interesting to consider the fact that in the indexing cases linking WM individuals to WM episodes, the linking happens mainly through the activation of *reafferent signals* indicating completion of SM operations (e.g. [Spec,AgrSP],[Spec,AgrOP],[Spec,VP])—

and in one special case through activation of a *consequent state* (a whole WM situation representation). In the case of whole actions that update the WM situation, the reafferent states are very obviously *WM situations*, or rather *aspects* of WM situations.

So the rough idea would be like this: we get into a WM situation, and this 'indexes'—i.e. activates—not an *atomic* episode but a *macro* episode: a sequence of 3 WM episodes stored in LTM. We do the first of these. This updates the WM situation: but we're now in a different situation from normal: the WM situation in-and-of-itself suggests various 'afforded' continuations, but there's now another, separate influence on the distribution of possible episodes: we're in the middle of activating a prepared *sequence* of episodes, and the second of these is now strongly activated. This scheme allows for the execution of sequential plans, but also, importantly, for the interruption of plans if other contingencies arise in the WM situation. Finally, let's say we get to the end of a prepared sequence of episodes. At this point, the WM situation is updated in a special extra way: not only do we update to the consequent state of the last component episode in the executed sequence, but we also update to the consequent state of the sequence *as a whole.* (In language, we have now *drunk the tea*, or *fixed the bike*, or whatever.) The consequent state of one macro-action can be the precondition for another, and so we can build macros that have macros as elements, and so on up.

An important question concerns how macro-actions interface with language. When I say *I fixed the puncture*, I actually did lots of things: but firstly, the word *fixed* identifies these things collectively, and secondly, the words *I* and *puncture* are like variables in the routine, in somewhat the same way as they are variables in an atomic motor action like grasping. My assumption is that the arguments of the macro-action also feature in particular places in the actions it's composed of. This means that macro-actions are really very similar to motor programmes in lots of ways. For instance, you probably have to attend to the relevant agent, patient as a prerequisite to executing the action. But then when you select and 'execute' a macro-action, what you're doing is executing a prepared sequence of other actions, stored in LTM, using some pre-established pattern of variable binding.

Clearly, the idea is that words like *fix*, *clean* etc denote macro-actions stored as prepared sequences stored somewhere in LTM in PFC. But the thing is miles more complicated than that: the type of the objects that feature in these macros heavily influence the actions that are done, as famously documented by Pustejovsky (1995). So: fixing a bike is very different from fixing a chair, and so on. I'm not sure how this productivity works.

9.3 Statements of causality as resulting from 'internal' observation of the WM-situation-update function

I've suggested that LF structures featuring the verb *cause* describe things going on in the WM-situation-update function. But this function is presumably invoked *at every sentence* *boundary*. So sentences can *describe* what normally happens *between sentences*. One possibility is that LF structure featuring the verb *cause* report what happens when the observer focusses his 'internal perception' *on the WM-situation-update function*. I'm already assuming that the observer can enter 'feeling perception mode', and inspect *processes* relating to his own neural representations in the same way that he might inspect *processes* happening in the world. It seems likely that updating the current WM situation involves a process: a sequence of active SM representations in some particular medium. Maybe LF structures featuring the verb *cause* describe sequences of SM operations that result from *perceiving* this internal neural process.

One interesting idea is that

9.4 An idea about forward and backward chaining in the cause-identification network

[I think cause identification has to involve looking at the difference between $P(e|c) - P(e|\neg c)$.]

Here's one possibility, based on what happens in feeling perception mode. In feeling perception mode, the observer activates a jumbled *set* of feelings, and this is the trigger that puts him into feeling peception mode. In this mode, a *single stimulus* is selected: namely, the one that's most activated (collectively) by the set of feelings. So information travels from the *set* of *pre-attentional* feelings to the *set* of candidate objects-of-feelings. At this point, a *single* candidate object-of-feeling is selected. And now we go back in the other direction: we see which *individual* feeling this *individual* stimulus causes.[4]

So maybe it's something similar in the cause-identification network. When there's an update to the WM situation, there are in fact *lots* of changes. We might say: *if there are enough changes*, we're going to activate the cause-identification network. The first operations in this network are pre-attentional. The *set* of changes to the WM situation are presented as input to the network, which computes (abductively) a *set* of candidate causative eventualities (events or states). From these, a single candidate causative eventuality is selected (which then becomes attended to), and it's associated (*de*ductively) with a *set* of possible effects. We might say: if this set *includes the observed change* (say the melting of the paint), then we can conclude that the

This doesn't feel quite right: but I think this might be because my account of abductive inference is not right.

Don't forget: the sentence *The heat caused the paint to melt* must also be understood as an occasion in which a causal relation between the heat and the paint-melting is *learned*. As in the cup-grabbing sentence: the SM routine indicates how the action is experienced by an observer who has *already* learned the necessary SM structures—but it also indicates the mechanism by which these structures were *originally* learned. In the current case, *The*

[4] Perhaps we're ensuring that associations are not made across the board linking arbitrary stimuli with arbitrar feelings: this kind of 'cross-situational' learning of associations would work in the long run, but it's not very efficient.

heat caused the paint to melt has to signal a learning episode, resulting from the 'successful use' of a given cognitive network. Maybe what's happening is: (i) we *hypothesise* that the heat could be a cause of the paint melting; (ii) we see if adding the heat *by itself* to the previous situation is enough to trigger a situation update in which the paint melts; (iii) if it *is*, then we learn a link *specifically* between the heat and the paint melting. (Maybe this is an extra kind of learning that contributes to the WM-situation-update function, or something like that.)

Some random thoughts:

- Maybe the 'successful use' of the cause-identification network involves obtaining information about conditional independence? Either from the world, or from the simulation of the world encoded in the WM-situation-update function?

- To put this another way: there's a well-known idea in theories of causal reasoning (and Bayes net theories in particular) that learning causal models involves a good deal of active intervention.

- A toy example: if [pop-music] causes [John-gets-angry] when it's [early-in-the-morning], we have to create a node [John-gets-angry] in the Bayes net which is linked to by *both* [pop-music] *and* [early-in-the-morning]. In this case, the [John-gets-angry] node will have to specify the conditional probability distribution for [John-gets-angry] for *all combinations* of [pop-music] and [early-in-the-morning].

- Something that might help think about the cause-identification network in Bayesian terms is that I'm directing it at the WM situation—which is explicitly a *distribution of possible episodes*. We know that Bayes' rule is best solved by representing it as applied to whole distributions, because this means we don't have to know the scaling factor that relates $P(\mathit{effect}|\mathit{cause})$ to $P(\mathit{cause}|\mathit{effect})$.

- I suppose it could be that the situation-update function is *nothing more than* a Bayes net: it could be the whole update function, rather than just a sub-function that *contributes* to updates. I actually like this idea. If there was something to the update function *other* than updates due to causation, they would be based on pure association—and we know association is not very reliable by itself. Also, a Bayes net that represents cause-effect relations is a very *economical* way of making predictions about future states.

- A useful reference relating to how causal relations are learned developmentally is Gopnik and Wellman (2012). The model proposed here is explicitly Bayesian. What's being learned (I think) are arcs in a Bayes net. Certainly Gopnik and Wellman's proposal emphasises the role of *experimental interventions* in the (active) acquisition of causal knowledge. And it references Waldmann *et al.*'s (2008) work as an example of 'recent developments' in models of causal reasoning.

- A complete aside: if I'm thinking that the cause-identification network is also operative in a representation of internal causative processes linking external stimuli and feelings, then somehow these episodes 'taking place in the brain' have to be represented as updates to the WM situation. So the WM situation has to include reference to: (i) the observer's internal state, and (ii) 'the presence of external stimuli'. I can see how (i) might feature, but (ii) seems tricky: the WM situation is a distribution of episodes, and stimuli seem too small. One possibility is that in feeling perception mode the WM situation ceases to represent the outside world altogether, and just represents the observer's internal milieu.

- In the model of Waldmann *et al.* (2008), causal relations are learned *one at a time.* Or more precisely, what's learned is a little local piece of a Bayes net, involving one effect e and one candidate cause c. The crucial data to get is the difference between $P(e|c)$ and $P(e|\neg c)$ (also called ΔP). This effectively sums over all other possible causes, both observed and unobserved, known and unknown. Actually in the general case, you have to learn the conditional probability table for *all* things which are hypothesised to be causally related to e: there could be several. (But I guess if you find them one by one, that's a start: having found them, you can start learning about their effects in combination.) Actually I think that in Waldmann *et al.*'s model, which they call a 'single-effect' learning model, and which they are at pains to distinguish from causal Bayes nets, the agent explicitly *doesn't* represent the full conditional probability table involving all causes.

One way of construing the cause-identification network in the framework of the WM-situation-update function is to assume that the network doesn't represent all of the nodes of a Bayes net simultaneously; it just represents projections from a single node in such a network (either back to possible causes, or forward to possible effects). In this scheme, a Bayes net is only implicitly represented. Doing extended Bayesian reasoning would probably involve some kind of recurrence, so that if we reason from the cause E_1 to the effect E_2, if we want to reason further, we have to copy E_2 into the cause layer.

9.5 Where else does the state-change detection mechanism feature?

It's interesting to think again about action recognition in the context of the perceptual state-change-detection mechanism. An animate agent is (among other things) an object that can change its physical state. When the agent's physical state is changing, that's (at least normally) when the agent is doing an action. The action can be parsed in two ways. Firstly it can be represented in the motor system. But secondly, as discussed at length in my account of grasping a cup, while the grabbing action is 'under way', the agent of the action is (re-)represented as a pattern of movement. Maybe this analysis of the agent as a pattern of movement happens in the same perceptual modality that represents the changes in an inanimate object (e.g. paint while it's melting).

Chapter 10

An extended model of WM episodes

This is currently all about verbs with three arguments; I'm sure it can be better structured.

10.1 Verbs with three arguments

10.1.1 Ditransitives and benefactives: a role for the interlocutor

Syntactic theories of argument structure have to take into account straight ditransitives, like Examples 10.1 and 10.2:

(10.1) John gave Mary the cup

(10.2) John sent Mary a letter

They also have to take into account **benefactives** like Example 10.3:

(10.3) John made Mary a cake

The beneficiary *Mary* is an example of a 'non-core' argument (also called an **adjunct**). It's not obligatory, either syntactically or semantically, but it denotes a participant that *can* feature in an episode.

Ditransitives and English benefactives are frequently analysed as deriving from a causative LF structure (see e.g. Harley, 2003). On this analysis, Example 10.1 realises the message 'John caused [Mary received the cup]', Example 10.2 realises *John caused [Mary received the letter]*, and Example 10.3 realises *John caused [Mary recieved the cake].* This analysis works well for Example 10.1, but for the other two examples a problem has been noticed (Pylkkänen, 2002): asserting that John sends Mary a cake does not imply that she receives the cake; likewise, asserting that John makes Mary a cake does not imply that she receives it. So it seems wrong to talk about causation in these sentences. Pylkkänen's alternative analysis is that in the semantics of these sentences, the beneficiary is not related to the episode, but merely to the created or transferred object, which is asserted as being transferred 'to the possession of' the beneficiary. But this has problems of its own. It is important that the possession is brought about as a result of the episode, rather than in some arbitrary way (Larson, 2010). If John makes a cake, and *Bill* gives it to Mary, it does not follow that John made Mary a cake. Pylkkänen's analysis of benefactives involving a simple possession predicate is too weak, and Harley's analysis involving causation is too strong, so some intermediate alternative must be sought.

An interesting solution is proposed by Beck and Johnson (2004). Their account draws on an account of the semantics of progressive sentences like Example 10.4.

(10.4) John was crossing the road.

The episode referred to in this sentence is one in which John crosses the road. But the sentence does not assert that this episode actually occurs; his action could be interrupted in various ways. The progressive auxiliary *be* (with associated verb morphology *-ing*) scopes the road-crossing episode within a modal operator, so that what is asserted is that John begins to cross the road and that in normal circumstances we would expect he would complete this action. Beck and Johnson propose that a similar modal operator is implicitly introduced in a ditransitive or benefactive. On this account, Examples 10.2 and 10.3 assert that John performs an action (sending or making) that begins a process which in normal circumstances we would expect to culminate in Mary receiving the letter or the cake. So what John's action causes is not Mary's receipt of the item itself, but the *initiation* of a process which can be expected to culminate in her receipt of the item.

In the account of Knott and Lee-Hand, a causative sentence like *John opened the door* is analysed as describing the execution of a sensorimotor routine involving a causative motor action. Assuming the observer is John, the observer decides to act, then attends to the door, then executes a special motor action 'cause', which engages a motor circuit that produces a movement which in the past has reliably led to the perception of the opening of the door. In the case of door opening, this perception actually occurs. But Knott and Lee-Hand's analysis doesn't actually assert that John's action caused the door to open. What's asserted is that John executes an action on the door which in his motor repertoire is reliably associated with its opening. Knott and Lee-Hand's account doesn't have to stipulate an implicit modal operator scoping the episode normally caused by the causative action: it's already part of the motor model. So it may give a nice treatment of the semantics of ditransitive and benefactive constructions.

However, Knott and Lee-Hand's account of causative motor actions only considered simple causatives where the manipulated object undergoes an intransitive (or more precisely, unaccusative) action—for instance bending, breaking, opening. In order to give a sensorimotor account of ditransitives, I must extend Knott and Lee-Hand's account of causative motor actions to include actions that cause target objects to move to goals.

10.1.2 A sensorimotor account of a ditransitive action: giving and throwing

I'll consider the ditransitive sentences *John gave Mary the ball* and *John threw Mary the ball*, because the episodes described by these sentences are maximally concrete. The main difference is that the former sentence requires that Mary gets the ball, while the latter does not: if things don't go as expected, she may not end up with the ball. (Throwing is also interesting in its own right, because it's something humans can do, but no other primates; see e.g. Westergaard *et al.*, 2000).

Taking Knott and Lee-Hand's model as a template: *John threw Mary the ball* describes a SM routine in which John establishes Mary as a recipient, and then executes a causative action referred to this recipient that in the past has reliably led to the perception of a certain episode involving the recipient. This causative action can also be analysed as the transitive action of throwing the ball.

In this case, the causative motor action (throwing) clearly has some internal structure: John must first physically reach/grasp the ball, and then when he has it, he must do the throwing action on it. Reaching-to-grasp the ball is something that's already been modelled: it involves attending to the ball to generate a goal motor state, and then acting so as to achieve this goal state. The new aspect of this reach action is that before it is carried out, John establishes Mary as a beneficiary—i.e. as some kind of 'joint agent' of the action. My first idea is that because John has established Mary as beneficiary, the deictic structure of John's throwing action ('attend-to-ball, throw') is somehow guaranteed to echo that of the experience to be caused in Mary. From Mary's perspective, she is receiving the ball: this involves first attending to the ball, and then receiving it. When Mary is established by John as a joint attender to his actions, any action of attention he does will result in her doing the same action. John naturally has to attend to the ball when throwing it. But in the mode where he's doing the action jointly with the observer Mary, he has to attend to the ball in a public enough way that Mary will be able to attend to it too. Then John has to throw the ball. Since Mary's attention is on the ball, she will experience the ball travelling to her—and he will represent this experience—while he is *also* having a motor experience, namely execution of the motor action of throwing. These two things don't actually overlap much: the motor throwing action more or less precedes the Mary-gets-the-ball episode. But John has to represent them both. [1]

Now consider *John gave Mary the cup.* In the episode reported by this sentence, John is giving the cup, and Mary is getting it. John has to experience two separate actions simultaneously. The key thing is that, after he's established Mary as experiencer, he only has to execute a single *attentional* action: he has to attend to the cup. Within his transitive motor programme network, he then has to execute 'give' on this object, and within his

[1] Recall that the whole idea of establishing someone as an experiencer is that once you've done that, the attentional actions you execute will be mirrored by the experiencer. In the simplest case, you are just getting the experiencer to follow your attention. For instance, in *John showed Mary the cup*, John makes himself the agent, then establishes Mary as experiencer, and then acts on the cup in a way that will make the cup salient to Mary. We can say John caused [Mary saw the cup], by performing an action on the cup himself, and this action was the action of 'showing' it. The showing action is just a transitive action. But it's combined with the action of engaging with Mary as experiencer. In simple physical terms, showing Mary the cup involves getting Mary's attention, then making the cup salient, while looking back at Mary, to emphasise the communicative nature of the action, and checking whether Mary is understanding it by looking to the cup as well. Think of Tomasello's 'communicative gestures', that involve alternation of attention between the shown object and the interlocutor.

perceptual network, at around the same time, or a later time, he has to perceive Mary 'getting' *this same object*.

10.1.2.1 Parameters of a ditransitive action

As a physical action, the 'give [Mary, cup]' action has *two* parameters. After deciding to act, John has to attend to two entities: first Mary, then the cup. Each action of attention defines a goal motor state: the first relating to the cup, the second to Mary). These goal motor states constitute the two parameters for the physical give action.[2] The two parameters are defined very differently. The first parameter is defined indirectly. When John selects Mary as the beneficiary, this places constraints on the second goal motor state: it will have to make some reference to Mary. The second goal state is not defined as 'a place near Mary': it's defined as a place *relative to Mary* which if established as the second goal motor state in the causative action reliably results in Mary receiving the cup. The second parameter is also probably defined indirectly. Having established Mary as beneficiary, John then attends to the cup as a goal motor state, so that he does some gesture on the cup (most obviously, grasps it, or perhaps just does some kind of 'showing' or 'presenting' action): again, whatever operation on the cup causes Mary to receive it. The first effect of this action will be that Mary *notices* the cup. She will then *take* it. So whatever John's causative 'give' action is, it must result in Mary taking the cup. This means John might *offer* Mary the cup, or *pass* it to her, or something along those lines.

10.1.2.2 A more precise SM model of giving Mary the cup

Preliminary assumption: the physical action involved in giving Mary a cup (i.e. the motor action that *causes* her to receive the cup) has two parameters, each of which is a goal motor state computed through an action of attention. These goal motor states have to be attained in sequence: the first relates to contact with the cup (in the simplest case, a stable grasp on the cup); the second relates to the location the cup should be transported to (in the simplest case, somewhere near Mary).

The action itself involves various perturbations of these goal motor states: John can't reach straight for the cup, and when he's got it, he can't reach straight for the goal location: in each case the hand must describe particular trajectories. The perturbations are specified by a motor programme—for instance 'give', 'throw', 'flick'. For now we're not concerned with exactly how these perturbations work; we're just concerned with the two parameters, i.e. the two goal motor states.

(1) John attends to himself, to select himself as the agent of a forthcoming action. Then (2) he attends to Mary as beneficiary. This has two effects. Firstly, it makes Mary a *joint agent* in John's forthcoming action. So everything in John's SM routine from here on in has to be sufficiently public that Mary can retrieve the routine herself, and play her part. Secondly, it places constraints on the second goal motor state for John's causative

[2] Notice that the motor states are in the opposite order to the original actions of attention, which is interesting.

action. This goal motor state has to be such that the action as a whole results in John perceiving Mary receiving the ball. In the simplest case, this means the goal motor state should be near Mary. In a more complex case it means the goal motor state should have some *relation* to Mary. The relevant properties of Mary are quite abstract: they include her direction and distance from John (since these determine the direction and force/speed of the throwing action, or whatever action results in John perceiving Mary receiving the ball). They may even include her address: if John is *sending* Mary the ball, then his causative action involves putting the ball in a parcel, writing Mary's address on it and posting it.

John then (3) attends to the ball and generates a default goal motor state associated with the ball. This is the first motor state in his motor action. This means he's going to do something on the ball—which in turn means that, now or later, Mary's attention will be drawn to the ball (since she's watching what he's doing).

John then (4) selects and executes a motor programme defined in relation to the two specified goal motor states. The motor programme will bring the hand/arm first to the first state (touching/holding the ball), then to the other (relating to Mary), and will do so by passing through various intermediate points ('perturbations', in my terminology). As a whole, the motor programme is represented in the causative actions network, as being the programme that in the past has been found to cause a particular episode to be perceived, namely Mary receiving the ball. And finally, (5) John perceives the episode of Mary receiving the ball. This caused episode is different from the caused episode of a lever bending or a door opening, because John has already established Mary as an observer, which means that the episode as John experiences it is also the episode as Mary experiences it: he's seeing it through her eyes, in some sense. (He doesn't experience the lever bending 'from the perspective of the lever'.)[3]

My main questions now concern the representation of the caused episode in the PFC-based planning medium. *John can't bring about the caused episode ['Mary receives the ball'] unless this episode is represented in the current planned sequence in PFC.* Recall that the causative actions network gets input from the PFC medium holding a planned SM sequence; this is what enables the desired episode to affect motor movements before the episode itself actually occurs. There are two questions.

10.1.2.2.1 (1) How does PFC represent a planned episode? There's still a sequence to be brought about: Mary must first attend to the ball, and then perceive herself receiving it. The operations 'attend-ball', 'get' would be the last two operations in the planned SM sequence, active in parallel, and therefore able to influence both phases of John's causative motor action in parallel.

[3] Does Mary experience the throwing part of the action as well? Mary could presumably notice that she receives the ball because John throws it to her. But I think that's not part of John's goal: he just cares that she gets the ball.

10.1.2.2.2 (2) How does John know to first attend to the ball? If the planned actions are active in parallel, and the causative actions network is responsible for determining the form of the movement, how does John know to attend to the ball *then* do the other movement? My idea here is that John's attention to the ball is also just a normal action of attention to the ball, of the kind he'd have to do in preparation for grabbing it. Part of what John is doing is a normal action on the ball. In a case like *John grabbed Mary a cup*, the verb emphasises the component of the action that's a completely normal grab action, [attend-to-cup, grab] (though the sentence also conveys a special intentional state that endures after the grab action is complete). My suggestion is that after attending to Mary as joint agent, one component of John's action is no different from normal: he still *performs an action* (in this case, a grab). The only difference is that he does it in order to achieve an effect on Mary, to make her experience something. (Getting the cup.) He's left in a state where he expects Mary to get the cup.

There's the causative action (the grabbing, plus something extra afterwards) and the caused action (the perception of Mary getting the cup). John *grabs* the cup, and Mary *gets* the cup. Both the causative action and the caused action have the cup as transitive object. My idea is that the caused action has the cup as transitive object *because* the causative action does: because John identifies Mary in advance as beneficiary/joint-agent, then executes CAUSE to make her a proper agent, when he attends to the cup as target, the action he will cause in her—perhaps some time later—is also the action of attention to the cup; and his physical action on the cup (first grabbing *then something else*, in this case some enduring intentional state, which will eventually result in his offering her the cup. It might be this latter thing which causes her to attend to the cup, but it's still ultimately attributable to his attentional action.

10.1.3 Verbs with a direct and an indirect object

Now consider the alternative argument structure for *give*: *John gave the cup to Mary*, which in the causative analysis is 'John does an action which causes [the cup goes to Mary]'. In fact, for concreteness' sake I'll consider *John put the cup on the table*, which in my analysis is 'John does an action which causes [the cup goes onto the table]'. In this case, John makes himself the agent, then attends to the cup, then [I think] decides to execute a causative action, which results in him attending to the cup 'as an agent'. (So now if there's any perspective being adopted, John is adopting the perspective of the cup.) In the kind of causative analysed by Knott and Lee-Hand, John just has to execute an action that causes the cup to undergo an unaccusative movement like breaking or tilting. But now there's an extra stage to the perceived movement of the cup (somewhat analogous to the extra stage of having to attend to the cup after establishing Mary as beneficiary in the earlier scenario). The thing that's intended is a locomotion action. To execute a locomotion action first requires selection of a locomotion goal, and then monitoring of a trajectory (and at the end possibly re-establishment of the locomotion goal as the new actual state). In this case, after having established the cup as an agent (i.e. having gained control of the cup as a manipulable object) John must attend to the goal location to which the cup should be moved, and make a physical movement, to move the cup along a particular trajectory—the key thing is that the attentional action and the selected movement should jointly cause the cup to actually move along the selected trajectory to the selected goal location.

In this latter case, I need to remember that movement trajectories are defined in allocentric coordinates, not in motor coordinates; what the agent is doing is causing the end-effector (whether it's his hand or a moved object) to locomote along a trajectory that makes sense perceptually, in allocentric terms). After John has attended to himself, attended to the cup and grasped the cup, he now has to do the action that causes the cup to describe a particular trajectory to a particular perceptual end-point.[4] The first thing he does is to attend to the goal location: an area on the surface of the table. This defines the goal location in perceptual terms. I suggest it also serves to identify a goal *motor state* of John's hand/arm. To recap: there are two SM routines being experienced alongside one another, that in this case overlap almost perfectly in time: one is John's causative motor movement; the other is the perceived travelling movement of the cup. John's motor movement has to achieve a particular motor state. Separately from this, the cup must be perceived to end up in a particular place. Importantly, the hand and the cup don't end up in the *same* place; the cup will end up on the table, but the hand doesn't have to. But I suggest that John's action of attention to the target location of the perceived action (a point on the table) will *also* serve to specify his goal hand state. If we were being simple, his goal is to get his hand to the perceived target location. But we can be a bit more subtle. At the time John attends to the target object, he has already established a grasp on the object,[5] so he has constrained the location and orientation of the object's base in relation to that of his hand. We need to envisage another function that learns *perturbations* of goal motor states here, I think. I assume that the surface of the table (the goal bit of surface) is initially represented in terms of its motor affordances (see Section 13.8 [**??**] of Part 2): specifically, as the motor state in which the hand is flat against the surface. The particular object that John is transporting—a cup—plus the grip that John has on it—e.g. holding it by the handle or rim—should be represented as a perturbation of this goal state: a certain degree of rotation and perhaps a slight change of grip. The trajectory relative to the goal state is again represented very differently for the causative motor movement and the perceived locomotion episode. This is partly a matter of different coordinate systems. But that's only part of it. The causative motor action can be very different: not in the case of 'putting', but in the case of 'throwing' or 'sending' and so on.

I'll briefly consider throwing. In this case, John's motor trajectory reaches a radically different goal motor state, which is defined in body-centred coordinates and also with reference to the current body-centred position of the goal location. (John can throw to the left, to the right, behind himself, while looking left, right, straight ahead, etc). The trajectory taken *to* this goal motor state is wildly perturbed; he has to draw his hand back

[4] I'm simplifying by assuming that the causative action only starts when John is grasping the cup. The desired effect also influences the approach trajectory of John's hand onto the cup. But forget about that for now.

[5] Or he will shortly have done so. The agent's gaze normally moves to the target location a little before contact is made—presumably at the point when the nature of the grasp can be well predicted.

away from it, and then move towards it like a slingshot. On the other hand the perceived trajectory of the ball is a parabola.

To summarise: in both *John gave Mary the cup* and *John put the cup on the table*, there are two actions. In each case, the first action is a physical action of John's, that has two goal motor states; the second action is an action of something other than John (Mary, or the cup), perceived (or at least intended) by John, that is *caused* by John's motor action. In the first case, the relationship between the two actions concerns a *joint action*, in which one agent engages collaboratively with another. In the second case, the relationship between the two actions turns on the way a motor action is controlled by a perceived action in a different coordinate system.

10.1.3.1 Locative clitics

Consider *Jean y allait* ('John went there') or *Jean y posa la tasse* ('John put the cup there') or *Jean l'y posa* ('John put it there') or *Jean me l'y posa* ('John put it there for me'). The goal location, if pronominal, can be expressed as a clitic, even when the direct object and/or benefactive is a clitic as well.

If clitics are read from PFC sequence plans at the same time as verbs, as I suggest, then the goal location must be part of the plan. John plans to attend-to-beneficiary, then to activate CAUSE, then to attend-to-cup, then to activate 'go', then to activate goal-location. These things are all active in parallel, so can be read out at the same time: e.g. we can read out the pronominal beneficiary, then cup, then goal-location, then cause-go (read out jointly as *put*) with inflection agreeing with the subject.

10.1.3.2 PP fronting for heavy object DPs

Consider *John put on the table [his wallet, his gun and his badge].* The object DP is heavy and therefore goes to the end. How would I analyse that? I don't think the PP *on the table* can be read out in full from the planning module. But I don't think there's a position at the bottom of the LF structure for the object DP. So it's the PP that has to appear higer. I think this idea is similar to the proposal of Brooke (2008).

The P by itself can certainly appear higher—as in phrasal verbs like *pick up*, *put down.* I think the P in these cases is a straight clitic, pronounced at the same time as the verb, from the planning medium.[6]

To raise the whole PP, one possibility is that you can use the planned trajectory of the target objects as an opportunity to *read out* the PP, in the same way that an argument DP position provides an opportunity to read out a WM individual. A PP is after all a unit

[6] That idea has a couple of consequences. Firstly, this kind of 'directional clitic' has to be distinguished from the locative clitics illustrated in French above. Both types of clitic have to be visible in the planning medium, but for some reason English allows directional clitics but not locative ones, while French does the opposite. This has to be explained. Secondly, if the P in a phrasal verb is a clitic in English, then object pronouns in these contexts are also clitics—as in *pick it up*, *put me down.* The only way they could get in front of the clitic P is if they are clitics themselves. So they would be the only object clitics in English.

in its own right, associated with its own coordinate system, like a DP. This is my current proposal.

How to ensure the PP is only read out early if the object DP is heavy? I think the heaviness of a DP could be added as a feature, and then this policy could be learned. As for why it evolved in the language in the first place, I think this is to do with communicability; I don't think I need to consider it, just treat the convention as a fait accompli.

10.1.4 High applicatives

English benefactives only work with creation verbs, as discussed above. In other languages, benefactives can be used with verbs more productively. The Bantu language Venda is an example of a very productive language (see Pylkkänen, 2002). Here unergative verbs like *speak* can take benefactives, conveying who the action was done 'for':

	Mukasa	o-amb-el-a	Katonga.
(10.5)	Mukasa	3sg.PAST-speak-APPL-FV	Katonga
	'Mukasa spoke for Katonga'		

Pylkkänen argues that the productive type of benefactive is syntactically and semantically quite different from the English one: she calls the English benefactive a 'low applicative' and the more productive benefactive a 'high applicative'.

French is an intermediate case: applicatives are more productive than in English, but less productive than in Venda (see e.g. Boneh and Nash, 2009). In French, unlike English, clitic indirect objects can express beneficiaries, even when there's no posession.

	Jean	lui	mangea	du pain.
(10.6)	Jean	for-him/her	ate	of bread
	'Jean ate bread for him/her'			

However, French applicatives can only work in this context when there is a direct object, even if this object isn't in any sense transferred to the experiencer. In Venda, there is no requirement for a direct object.

	*Jean	lui	mangea.
(10.7)	Jean	to-him/her	ate
	'Jean ate for him/her'		

But I would prefer to see them as occupying the same high position, relating to John's establishment of him as an experiencer: John is doing the action *so that it is experienced by him.*

I suggest the difference between high and low applicatives is really to do with whether there's a causative or not. In the high applicative *Mukasa spoke Katonga* ('Mukasa spoke for Katonga'), there's no causative: Mukasa is speaking, and presenting this action to Katonga as an experiencer (whose goals are taken into account). In the low applicative *Mukasa laughed Katonga* ('Mukasa made Katonga laugh'), Mukasa is establishing Katonga as an experiencer, and then doing a causative action that results in Katonga doing something (namely laughing), in much the same way as in *John gave Mary a cup* John establishes Mary as an experiencer, and then does something (a giving action) that results in Mary doing something (namely receiving the cup).

I like the idea that you can't do an action for the benefit of an experiencer unless it's an action done *on an object* of some kind, because the experiencer has to be able to establish *joint attention* on the object as part of sharing the action. [In that case what's going on in Venda *I ran him*?]

Verbs of creation can take an adjunct specifying the agent for whom the object is created: for instance *I made her a cake*. This adjunct must denote a sentient agent. One possibility is that it is represented within the interlocutor medium. The action described by the verb is in some sense communicative, as its purpose is for the sentient agent in question to perceive it (or its effect). If I make Mary a cake, but Mary doesn't eventually get the cake, the action didn't go as planned.

10.1.5 Coreferential benefactives

Boneh and Nash (2009) discuss another type of high applicative in French, illustrated in *Je me suis maté un film*, translated as the Appalachian *I watched me a film*. This is an applicative obligatorily coindexed with the subject. Boneh and Nash argue persuasively that it's different from a normal benefactive. (In Appalachian, there's a semantic difference between *I washed me some clothes*, and *I washed myself some clothes*: the latter sentence means something like 'I caused (through washing) myself to receive some clothes', whereas the former one doesn't convey any transfer of possession of the clothes to the agent.) I like the idea that these coreferential datives are still benefactives: what the agent is giving him/herself in this case is the pleasure of *being the agent of the action*. You can't give *someone else* this pleasure, obviously, which would explain why these benefactives have to corefer with the subject. More precisely, what you give yourself is an agent's-eye experience of the *object* of the action, rather than of the whole action. When I drink me some beer, I'm causing myself to experience *the beer*, through the drinking action, rather than to experience the whole drinking action; the drinking action is just the causative action which enables me to experience the beer. Drinking is an *attentional* action as well as a substantive action. In *I drank me some beer*, what's being emphasised is that when an agent drinks beer he *perceives* the beer in a special way: you can only *taste* the beer when you drink it. (Taste is the equivalent of 'haptic establishment' for grabbing: you experience the beer first visually, or haptically, and then haptically.) Receiving the *experience* of the beer (through a drinking action) is completely different from receiving the beer as a physical object (through a giving action). This analysis would explain why coreferential applicatives require direct objects. And also why they're much better with indefinite objects. A definite object is one that has already been experienced.

Even in *I shot me a pig* we can argue the pleasurable experience is of the pig rather than of the shooting. According to my SM analysis of transitives, the agent of *I shot a pig* experiences the pig twice: first through a passive perceptual process, and second (in the consequent state of the shooting action) 'in a different modality'. Obviously this new modality is not a new sensory modality in this case; there's no *tactile* experience of the pig, unlike in the action of grabbing (or drinking, for that matter). But the pig is definitely in a different *state* afterwards, and it's by witnessing actions involving changes of state that an observer can learn the different states that objects can be in. So not all learning linking high and low object positions relates to cross-modal object representations; some of it relates to learning about the different states that objects can be in. When you've shot the pig it's still a pig. In any case, I'm arguing that it's not the action of shooting that's pleasurable: it's the experience of watching the change of state on the pig that my shooting action brings about. (At least this is pleasurable for the kind of person who would say *I shot me a pig*!). So conceivably what the agent of the action is doing is causing himself—through the action of shooting—to have this experience.

In all these cases, we have to envisage a special sensorimotor operation through which an agent preparing to act attends to himself as the recipient of a forthcoming experience. John prepares to drink, but additionally attends to himself as an experiencer, so he will actively experience the beer while drinking; John prepares to shoot, but additionally attends to himself as an experiencer, so he will actively experience the change of state on the pig that his action brings about. I like the idea that John is causing an experience in an observer or joint agent through his physical action, just as in the case where John is giving Mary a cup. I actually like the idea that he's establishing himself as an experiencer of his own action. But what actually changes cognitively when an agent does this? The agent is *watching himself*, so we should expect there to be *some* change.

I suggest that this sensorimotor process is very similar to that described by a straightforward experience verb, for instance *John saw the cup*, or *John tasted the beer*. ('Tasting the beer' is very different from 'drinking the beer'—it focusses on the sensory experience rather than on a substantive action.) To recap from Part 2 Section **??**: in the process described by an experience verb, we first make ourselves the agent, and we then 'non-standardly' go into perception mode. Since we do this *as agents*, the operation of going into perception mode is registered as an action. (Normally it's associated with a semantic feature of the AgrS head; nothing to do with the V head at all.) In the case of *I drank me some beer*, the agent is doing a real action, but *also* establishing perception mode, because he wants to perceive something relating to *his own action*. We have exactly the same sequence [attend-to-agent, establish-perception-mode] as in an experience verb. However, the establishment of perception mode is not represented as a *verb* (e.g. *see*, *taste*), but as a *participant* (namely the optional experiencer participant). This is a neat way to render the agent's experience of his own action as well as the action itself. The substantive action (e.g. drinking, shooting) is expressed by the main verb; the experience verb is hidden.

So: is there implicit causation in a coreferential benefactive? Does *I drank me a beer* mean 'by drinking the beer I *caused* myself to experience (by tasting) the beer'? Or is it analogous to Venda *I ran John the race*, which means 'I ran the race *in service of a goal of John's*'—i.e. 'I drank the beer *in service of a goal of my own*'? I think we must assume causation. I drink the beer *because* of the experience it will cause in me. Also if I'm just saying the action was undertaken in service of one of my own goals, I'm not really saying anything new; and I'm not saying anything about myself as experiencer, rather than just as agent; and I'm not explaining why there has to be a transitive object.

10.1.6 Unaccusative causatives and voice-bundling

In Japanese, there's a construction in which there's a causative verb (e.g. 'cause-to-die') with the patient on which the cause acted, but in which the causing action/event is implicit. This is shown in reading (ii) of the following example.

(10.8)
Taroo-ga musuko-o sin-ase-ta.
Taro-NOM son-ACC die-CAUSE-PAST
(i) 'Taro caused his son to die'
(ii) 'Taro's son died on him'

The example and glosses are from Pylkkänen (2002); I'd prefer to gloss Reading (ii) as 'Taro, [something] caused-to-die the son'. Pylkkänen argues that the implicit subject of the causative verb has to be a whole event, rather than an individual.

In Finnish, there's a similar causative structure with an implicit causing event, in the so-called **desiderative** construction.[7]

(10.9)
Maija-a naura-tta-a.
Maija-PART laugh-CAUSE-3SG
'Maija feels like laughing'

I'd prefer to gloss this as '[Something] is causing-to-laugh Maija'.[8] (Or more idiomatically, '[Something] is giving Maia the giggles.') Again, the something has to be an event rather than an individual. Pylkkänen's analysis of the causative here actually suffers from the same difficulty she noted for analyses of 'give' as 'cause to get' (see Section 10.1.1): Maija doesn't *actually* have to laugh, but she will do so in the normal course of events, unless she actively checks herself. So there is no actual causation here.[9] An important adddendum to the above example is that the agreement affix on 'cause-to-feel' isn't agreeing with the partitive noun *Maija*; it relates to the implicit causation event (which is also 3rd person singular). This can be seen in another example, from Nelson (1998):

(10.10)
Minu-a laula-tta-a.
I-PART sing-CAUSE-3SG
'I feel like singing'

[7] Cathcart, 2011 calls these constructions 'impulsatives', and discusses Albanian examples too. Her generalisation is that impulsatives are best glossed in English with 'X feels like Ying', their subject has experiencer case, and their verb carries morphology that's non-active and doesn't agree with the subject.

[8] The progressive aspect of the gloss is intended to convey the role of the Finnish partitive case in conveying aspectual information—specifically some kind of unboundedness (see e.g. Kiparsky, 1998).

[9] It seems to me that the desire to laugh is actually nonvolitional: it arises through involuntary processes. Pylkkänen is certainly at pains to distinguish the desiderative construction from a regular desire predicate like *I want to laugh*, that in Finnish doesn't make any reference to causality.

In these languages, the main verb of a sentence apparently isn't required to have an explicit external argument—though the external argument is there implicitly, as it can be referred to in e.g. sluicing constructions ('Something caused-to-laugh Maia, *but she didn't know what*[10].') Pylkkänen calls constructions where the verb has explicitly causative morphology, but no external argument, **unaccusative causatives**. This is by analogy with unaccusatives like *The door opened*, which means '[Something] caused-to-open the door'.[11] Obviously in English we can't say *I laughed* meaning '[Something] caused-to-laugh me'. There's a parameter of variation between English and Finnish.

10.1.6.1 The voice-bundling parameter

Pylkkänen suggests this parameter concerns whether the projection introducing a clause's external argument (VoiceP) and the projection introducing a CAUSE predicate (CauseP) are independent of one another, or whether they are **bundled** into a single projection. She argues they are independent in Finnish and Japanese, which allow causatives *with* external arguments (as in 10.11 below) and causatives *without* external arguments (as in 10.12), as well as non-causatives with external arguments (as in 10.13).

(10.11) $[_{VoiceP}\ [_{CauseP}\ [_{vP} \ldots]]]$

(10.12) $[_{CauseP}\ [_{vP} \ldots]]$

(10.13) $[_{VoiceP}\ [_{vP} \ldots]]$

In English, she argues CauseP is bundled with VoiceP, so that the projection introducing a causing event obligatorily also introduces the agent of this event. So for instance, *Mary broke the glass* has the following structure:

(10.14) $[_{VoiceP}$ Mary $[_{Voice'}\ [_{Cause,Voice}$ [the glass break]]]]

In this structure, the head of VoiceP (the projection introducing the external argument) also introduces CAUSE.[12]

10.1.6.2 A sensorimotor interpretation of the Voice-bundling parameter

It's interesting that Finnish desideratives require the nominal argument to take partitive case, which is also the case assigned to the subjects of experiencer verbs. For instance:

[10] Cathcart (2011) says this sentence isn't accepted by all Finnish speakers, actually.

[11] Note this analysis suggests there's an implicit cause in this simple action, unlike in my proposal with Jeremy.

[12] I know that English doesn't have unaccusative causatives, but the argument structure of words like *get/receive* is worth thinking about. In *Mary got/received a cup*, we have to posit a sending event that Mary participates in as recipient, even though this doesn't surface in the overt syntax. This sending event could just be a semantic presupposition of *get/receive*, but to me it's interesting that Pylkkänen's cause-to-laugh unaccusative can be glossed as *I got the giggles*. Which arguably means 'Something gave me the giggles', and maybe even 'Something caused me to get the giggles'.

Minu-a pelo-tt-a
(10.15) I.PART fear-CAUS-3SG
'I'm afraid'/'Something frightens me'

Sands and Campbell (2001)

Cathcart (2011) argues that all so-called 'impulsatives' have experiencer case.

I want to think about the process of experiencing onesself getting the giggles. What is attended to? We become aware of a process taking place in ourselves, which in the normal course of events would result in our laughing. We can certainly suppress this causal process, so that we don't laugh. To me, the crucial thing is that this process takes place in *ourselves*: it's something that we perceive *as experiencers* rather than as perceivers of external stimuli. My guess is that the reason these sentences don't need a normal subject in Finnish is that the process through which the agent attends to himself as an experiencer *is the causative process.*

I want to contrast three situations. In one, the observer's attention is drawn to an external object, and he goes into action perception mode. Here, the subject is not 1st-person at all (it's 2nd or 3rd person). In another, the observer decides to make himself the agent, and then executes an attentional action (resulting in experience verbs like *see*, see Section 10.1.5). Here we get a 1st-person subject, and the perceptual process is conveyed by the experience verb. In another, the observer's attention is drawn to a process taking place in himself (e.g. the process that leads to his wanting to laugh). This can be thought of in two different ways. Firstly, we can think of it as an operation in which the observer attends to himself *as an agent* (an operation which would be expressed linguistically with a 1st-person subject), and then separately *decides to attend to a feeling* rather than to do an action. (This way, we get *John felt...* and a continuation like *a song coming on*, *a laugh welling up*, *pain welling up*, *a pain in his toe* and so on.) Secondly, we can think of it as an operation in which the observer establishes himself *directly* as an experiencer. (This could be conveyed directly with experiencer case on the subject, as in an impulsative sentence.) My idea is that this is not available in a language like English, because it doesn't have overt experiencer morphology.

The interesting thing about attending to a process taking place in onesself is that the attentional action can be *identified with* the process. When we attend to an external stimulus, the action of attention is clearly distinct from the stimulus. (For instance, the stimulus is a cup, and the action of attention is a direction of visual attention to a certain point in space.) When we attend to an internal process, there's no such distinction. The process is already an operation.

Take an example: the process which results in me feeling a song coming on (i.e. the process which will lead to me singing unless it's checked). Becoming aware of this process involves *interoception.* A lot is now known about how interoception works: in a standard case like feelings of heat or pain in the body, these stimuli are first represented in posterior insula, where they serve to regulate the autonomic system, but are re-represented in mid-insula and anterior insula (Craig, 2009); subjective feelings of heat and pain correlate with activity in anterior insula. The anterior insula is also activated when a subject smiles, or experiences joy, or has an orgasm (see again Craig, 2009), i.e. when feelings come on. There's a growing consensus that the anterior insula is involved in representing *attention* to interoceptive representations whose role in regulating the autonomic system is normally preattentional. For instance Craig (2009) sees the anterior insula as representing 'salient' interoceptive stimuli (i.e. interoceptive stimuli which are competing for attention); Critchley *et al.* (2004) argue that the insula as a whole implements the mechanism by which we can become aware of internal bodily processes, and anterior insula represents those bodily processes that we are aware of; Damasio (1999) also talks about pre-attentional interoceptive representations (which he calls the 'proto-self') and post-attentional interoceptive representations (that he calls the 'core self', I think). In summary: the process that results in me feeling the 'coming-on' or 'welling-up' of a feeling, or a song, or a smile, or a laugh, is the process of selecting a single pre-attentional interoceptive representation (competing amongst other pre-attentional interoceptive representations) for attention. Note, the feeling of an action coming on is in one sense perceptual and in another sense motor: Craig (2009) suggests that the motor component is implemented in anterior cingulate cortex, where activity also correlates with subjective feelings (like anterior insula) but which also has a role in action initiation.

Importantly, all these authors (Craig, Critchley, Damasio) also see the anterior insula as having a role in the representation of the self: specifically, the self that *feels*. (...)

And now we can bring the above two ideas together. Recall: we're thinking about Finnish sentences that can be glossed as '[Something] me-PART caused-to-laugh'. The weird thing about such sentences is that they have no explicit representation of the causing event. I want to argue that the process that results in me feeling a laugh coming on is also the process that establishes me as an experiencer. So I can express it linguistically by referring to myself as an experiencer, rather than by referring to the causative stimulus directly. I think my idea is that there's an XP headed by the CAUSE operation, whose Spec is an opportunity to express the experiencer *as an experiencer*.

Something to connect to: Craig's (2009) model of how 'feelings' (e.g. of cold, pain) experienced within the body are represented in the anterior insula. (Note Pylkkänen's gloss of the unaccusative causative: 'I feel like singing/laughing'.) See also Damasio's (1999) somatic marker hypothesis. Craig (2009) says that the anterior insula integrates feelings about all different aspects of the body, including feelings originating in the brain (which is an important part of the body, metabolically), which would include feelings associated with onset of involuntary actions like laughing, crying, some types of singing...

Note that the idea that an agent can *check* the process which will lead to him laughing, singing etc may be compatible with the model of Brass and Haggard (2010), in which the anterior insula is involved in a decision to 'veto' a forthcoming action.

Question: can I sustain this analysis for the Japanese adversity causative? To-Taroo [something] the-son caused-to-die. I think so. The whole thing about this one is that it requires an experiencer: Taroo experiences the causing event in himself, somehow.

I'll discuss the whole question of how feelings are expressed in language in Chapter 11 and the question of how casual relations are represented in Chapter 9.

Chapter 11

Language for expressing feelings

11.1 Survey

There are various ways of conveying a feeling in a sentence. The agent in all cases is going to be John.

11.1.0.2.1 Localised feelings of pain I'll start by thinking about feelings of pain, which are quite concrete, since they are localised. There are several dimensions of variation. The subject can be John, or the body part with the pain, or the pain itself.

One way to express a feeling of pain is to *John felt sore [in his leg]*. One is *John's leg felt sore.*

I can also say *John had*

The pain can be expressed as a nominal (*pain*) an adjective (*sore*)

11.1.0.2.2 Pains arising A pain welled up in John. John felt a pain coming on. (Or a headache.)

Note that emotions (e.g. anger, happiness) can also well up in John. We can use the term **feeling** to cover both cases.

In this case, the verb (*well up/come on*) describes the *arrival* of the feeling. And this arrival process has to be localised *in* an experiencer (*in John*). (My feeling is that this 'in John' is what's conveyed in a subject with experiencer case in Finnish, Albanian, Japanese.)

It's interesting that a laugh or a song are also nominalised actions. Other cases are *a picture* (an artist can probably feel a picture coming on) or *a square.* There's no special feeling associated with squares—not emotional enough, I guess—but a square is something you *make* with a particular motor programme; it's a good case of a nominalised action. As discussed in the 'Verbs of Creation' section of Part 2, when you make, or draw, *a square*, *make* just means 'execute', and *a square* is a reference to a structured motor programme that has a certain perceptual result. I think songs and laughs are the same: singing can be described as *X-ing a song* (*X* =*sing*, *belt out* etc) and laughing as *X-ing a laugh* (*X* =*give*, sometimes *laugh*)

11.1.0.2.3 Subjects of experience verbs The subject of *feel* can be John, or John's body part—but it can also be the object being felt. This is an interesting alternation.

(11.1) John felt sore.

(11.2) John's hand felt sore.

(11.3) John felt the cup.

(11.4) The cup felt cold.

Does *The cup felt cold* mean 'The cup caused John to feel cold [in the hand that feels the cup]?'

11.1.0.2.4 Towards an analysis of raising verbs The above discussion is a good stepping stone towards raising verbs.

(11.5) The cup seemed cold. (To John)

(11.6) The cup seemed [to be blue]. (To John)

The implication here is that *seeming* relates to a feeling that an experiencer has when perceiving some arbitrary event.

What's the feeling described by *seem*? A very strong one, I think: it's to do with *uncertainty*. And note that the anterior insula is *heavily* involved in representing uncertainty, or degree of confidence: see e.g. Singer *et al.* (2009) and many other papers.

Now: what about a case where the verb *seem* introduces a whole complement clause?

(11.7) *It seemed that* [the cup was blue].

I presume the analysis of 'seeming' is the same: what changes is the lack of focus for 'seeming'. The feeling is no longer localised to the cup. I suggest that instead it's localised—explcitly or implicitly—to the experiencer. It's quite possible for an experiencer to have non-localised feelings. We can say *John felt cold*, but also *It felt cold [to John]*. The thing to figure out is how the transition signalled by *that* works in this case. My idea is that *it seemed* is actually a reference to John's feeling by itself—

11.1.0.2.5 Propositional attitude verbs and feelings The difference between fearing, hoping and believing is also at least partly to do with the feelings that accompany the evocation of a certain propositional state. *Knowing* is also a feeling! A feeling of *certainty*—see again Singer *et al.* (2009). In these cases there's no possibility of raising, because the owner of the propositional attitude occupies subject position. But why should the feeling be expressed first, and then its content? I'm guessing that hoping is also an action that evokes a certain type of content (like wanting). I can hope to run, or fear to run, in the same way I can want to run—these are subject control verbs—or I can believe John to run—an object control verb.[1] (Then there's *I consider John foolish*, which is something else again.)

There's also *It frightens/worries/pleases me [that P]*.

11.1.0.2.6 Emotion verbs: ***Like*** and ***please***

A lot of the material in this section is more tidily expressed in Section 11.2. I've put the bits that contribute new things in bold font.

In English we say *I like Bill* (or *I like [that P]*). In Romance we say *Bill pleases me* (or *It pleases me [that P]*): we can do these in English too, though they're a little marked. *Like* seems to express a feeling as a verb: aspectually it describes something like an activity (similar to *I stroke Bill*). I suggest it's like a motor programme: I attend to myself, then attend to Bill, and then *perform the operation 'like'*, in the same way as I could perform the operation 'hit'. In each case I reattend to myself (as an emotion-experiencing agent) and to Bill (as an emotional association).[2] **The main difference between *like* and *hit* is that *like* is atelic: there's no natural end state. This suggests that you can continue to experience yourself liking X for an arbitrary amount of time. While you're doing it, the SM state is analogous to that in which you're iterating on an action (e.g. clapping your hands), or pursuing some activity (e.g. drinking water).**

A key thing about liking is that it has to be related in very different ways to two individuals: the experiencer and the stimulus. In this sense it's quite a lot like a transitive action. Liking episodes are opportunities to create cross-modal representations of objects that involve *emotional/affective* representations; motor action episodes are opportunities to create cross-modal representations of objects that involve *motor* representations.

Now consider *Chocolate pleases John*, where John is an experiencer rather than an agent. Since this is a regular tranisitive verb in some ways, there should be cross-modal associations described here too. I think the information structure of this sentence is important: John is in some sense already given at the start of the sentence. What John is doing is evoking the concept 'chocolate' and seeing what emotion is evoked by association. The verb *please* is causative—in French you can explicitly have *faire plaisir*—so the struc-

[1] I think the predicate has to have a stative interpretation here.

[2] **An emotional feeling like 'liking' is a completely different feeling from a physical feeling like pain. Pain is located in a part of the body: the agent as a whole can feel it, but always in a particular physical place. Liking is also referred to an object, but the object can be *any object*, it doesn't have to be a body part (and it's normally not a body part); more importantly, the feeling of liking is *about* an object, not 'in' it, and it's *not* localised to any particular place in the agent. It's the *whole agent* that experiences likes, dislikes, etc. (Note that emotional feelings can be felt in the heart or body, but that's precisely when they manifest themselves as *physical* feelings.)**

ture has something in common with the Finnish ‘Something caused-to-sing John’. The evocation of an emotion can be thought of as an action done by the experiencer, or as an ‘action’ ‘done’ *by the stimulus*.

In what sense can a stimulus do an action? My idea is that it can cause a feeling to ‘well up’. Importantly, the welling-up action *is not the feeling itself*. I suggest it’s the process whereby the feeling becomes *attended to*: literally, the process that takes place in the anterior insula, which is the competitive medium in which feelings are attentionally selected. There are two components to this process. Firstly, one feeling emerges as a strong winner in this competitive medium. Secondly, the whole medium of feelings is attentionally selected rather than some other medium (e.g. visual attention). (As I’ve discussed for various other types of inter-media competition.)

It is helpful to envisage different roles for these two components. Perhaps nouns that denote feelings (e.g. *pleasure*, *sadness*, *anger*) pick out specific representations in the anterior insula itself, while action verbs like *hurt*, *please*, *soothe* describe the processes whereby these same representations become active; likewise **perhaps the process of selecting the anterior insula medium as a whole is denoted by an action verb like *feel*, or by a DP including experiencer case.**[3]

What does it mean to say that a stimulus *causes* a feeling to well up, emerge, arise? (I’ll use the term **well up** henceforth.) A stimulus is not a person, so I presume there’s no question of ‘volitional’ cause. However, it’s useful to recall that volitional agents are often in the business of causing *experiences* in other agents (joint agents)—see especially Section 10.1.1—and experiences can be feelings. The concept of experiencer comes up in two places: firstly the joint agent is an experiencer, and secondly the self can be an experiencer. These two topics are connected: that’s the whole thrust of Pylkkänen’s (2002) thesis.

The concept of causality is mental. This idea stems from the empiricist philosophers, especially Hume (1777/1975): the suggestion is that causation requires ‘constant connection’ between two stimuli (and various other conditions). In neural terms, causality in this sense could be defined by reference to a process whereby a mental representation reliably causes a particular ‘experience’. Idea: that experience could be a sensorimotor routine (e.g. perception of an action)—or it could be the welling up of a feeling. In either case it’s the *activation* of a representation.

In Jeremy’s model of volitional actions, a causative action is an action done by an agent that reliably (after training) brings about a particular perceptual experience (activation of an action representation). We know the action is done by the agent because we attend to the agent first, and the action is the next thing that occurs. Recall: the action is *deictically referred* to the previously-attended agent: we know this is the agent of the action because we attended to this agent just before. Analogously, I suggest we *attend* to a stimulus: after this, any welling-up of a feeling must be attributable deictically to

[3] This really does feel a bit like Marantz’ idea that there are category-neutral roots. In my terminology, this would mean that (e.g.) both the DP system and the clause system have links to the representations in the anterior insula, and to the representations that *select* the anterior insula as a whole medium: so these representations can be pronounced as nouns or verbs.

the just-attended stimulus. This is an interesting idea: we have to think deictically about stimulus-experiencer situations, just as we do about physical actions and events.

Is there any difference between attending to an object and attending to a stimulus? I think that you have to be in different states. To attend to an object, you have to have made yourself the agent, or made yourself a perceiver. When an agent attends to an object, he does not attentionally evoke feelings related to the object. When a perceiver attends to an object (in the null context), he doesn't represent herself at all beforehand. I'm envisaging a special operation through which an observer establishes himself as an experiencer, which allows him to attend to a stimulus, and to the feeling associated with it.

The operation that establishes an observer as an experiencer is the one that selects the anterior insula as the active medium, in a competition between alternative cognitive media. The alternatives are things like perception mode and action execution mode. So maybe what's selected is not just the insula, but a *mode of sensorimotor processing*, in which the insula plays a part. Let's call this **internal perception mode**, and regular perception mode **external perception mode**.

To help think about what might happen in this mode: the big question is, how does being in internal perception mode allow one to attend to objects *as stimuli* rather than as normal objects? Well: when attending to a piece of chocolate, I guess in internal perception mode we want to inhibit the activation of motor affordances related to the chocolate: these are completely irrelevant to the chocolate as a stimulus. We also are uninterested in any movements that the object is making. Or rather: we're interested in movements that elicit feelings, rather than movements that help identify e.g. a category of action. I think internal perception mode certainly features associations between object representations (either types or tokens) and representations of feelings. I think these associations have some directionality: the stimulus must surely cause the feeling, rather than the other way round. So now we have two partial orderings. John has to establish himself as an experiencer before he can activate feelings. And in some sense, John has to activate a representation of a stimulus before he activates the associated feeling: because it's the stimulus that evokes the feeling, rather than the other way round.

An idea: John has to attend to a *physical object*, in a non-experience context, in order to *get* a feeling in the first place. (A pre-attentional one, maybe, at this point.) Getting a feeling involves attending to himself as an experiencer: these two processes are coextensive. Only after John has attended to himself can he attend to the feeling post-attentionally. Again, I think there's a really interesting deictic routine, involving attention and re-attention to both the experiencer and the stimulus.

We have to envisage that we start in the neutral 'initial context': we haven't decided what mode we are going to establish yet. It could be action execution mode, or external perception mode, or internal perception mode. Maybe to get into internal perception mode we have to attend to a physical object—and this is why the object is the grammatical subject (e.g. *The chocolate tasted disgusting [to John]* or *The chocolate disgusted John*). In a case like this, the agent has to do quite a lot to even *enter* internal perception mode: he has to put the chocolate in his mouth.[4] This is likely to be something that is apprehended in a previous SM routine, resulting in a discourse like *John took a bite of the chocolate. It disgusted him.*

I'll talk more about this scenario below, in Section 11.2.

11.1.0.2.7 Aside: experience of generic stimuli Is there anything strange with the discourse below?

(11.8) John took a bite of the chocolate. It disgusted him.

Experience verbs can definitely report generics (e.g. *Chocolate disgusted John.*) But can they report disgust about a specific object (e.g. this particular piece of chocolate)?

Note that even when used with generic objects, an experience verb seems to report a specific episode in which the experiencer *has* the relevant feeling. If I say *I like you*, or *You enthrall me* (not interpreted as a volitional activity on your part), or *Chocolate disgusts me* (with a bare nominal subject), I'm not describing a specific experience with an object in the world at a given moment in time. I'm describing a SM experience that occurs at a single moment in time, but it's an experience in which I evoke an individual-level representation of you rather than a stage-level representation: *this* is the representation which elicits the feeling. It still an experience that occurs at one moment in time—I can say *I didn't like chocolate yesterday*—but the 'chocolate' that I'm referring to isn't the chocolate-stage of yesterday, it was chocolate in general; *yesterday* refers to the time at which the disgust experience was evoked.

This means we have to think of deictic routines initiated when a generic or individual-level concept simply pops into the experiencer's attention. (When does that kind of thing happen?) And we have to contrast these routines with routines initiated when some specific or stage-level stimulus is attended to by an experiencer.

11.1.0.2.8 Unknown stimuli There are certainly cases where the experiencer doesn't know what caused a feeling to appear. These are cases that are expressed with unaccusative causatives in Finnish ('[Something] made-sing John'), and in English, either appearance verbs predicated of feelings/nominalised impulsive actions (*Pain/a song welled up in John*), or clauses with John as subject and the verb *feel* (*John felt pain / a song coming on*), or *get*-verbs (*John got the giggles*).

11.1.0.2.9 Stimuli that evoke actions in an experiencer? Note it's not just emotions that a stimulus can engender in an experiencer. **Chocolate can *please John*, or *make John happy*[5] but it can also *make John laugh.* It can't make him do any**

[4] I don't think this any more. The chocolate can compete as an emotional stimulus in an emotional saliency map, and if it wins, it's just the experiencer's feeling towards this stimulus that counts.

[5] Pleasing and making-happy are not at all the same thing: and what's conveyed by *You please me* in French is very different from what's conveyed by *You please me* in English. There's a lexical gap for the verb *plaire* in English.

action; only 'impulsive' actions (in the terminology of Cathcart, 2011) that are at least somewhat involuntary responses conveying feelings. (I say *somewhat* involuntary, because these involuntary responses can be behaviourally *suppressed*—a function to which the anterior insula certainly contributes, see again Brass and Haggard, 2010.) Humans are special in being able to suppress involuntary responses of this kind. (I think.)

11.2 Experiencer perspective

In this section I discuss in more detail the case above, where John attends to an object, and then experiences a feeling about it.

11.2.0.3 The stimulus: is it an external entity, or an internal representation?

When I say *John annoys me*—in its non-volitional interpretation—does that mean that John himself generates the feeling of annoyance in me, or that my cognitive representation of John generates this feeling? Internally, I imagine the mechanism is that the cognitive representation of the individual John is associated with the feeling of annoyance (e.g. in the insula). My idea is that in internal perception mode, an agent represents the onset of his own feelings: for instance, the verb *well up* or *come on* can represent the 'action' of a feeling arriving—as in *John felt pain coming on*, or *Frustration welled up in John.* So a feeling like annoyance or frustration or pain is definitely a sensation in the agent, rather than something in the world. My question is: what about the *cause* of a feeling? Is that in the world, or is it a cognitive representation of something in the world?

I can certainly say *The chocolate annoyed me*: this describes a feeling of annoyance that stems from some particular chocolate. But I can also say *Chocolate annoys me*: one possibility for this generic is that the agent activates the concept 'chocolate' in his head, and *this* is what activates the feeling, rather than any actual piece of chocolate. So **possibly in *The chocolate annoyed me*, we are stepping from attention to an object in the real world *into the observer's experience*—crossing the boundary from perception mode into experience mode—and thereafter talking about an association between a concept and a feeling.**

Think about it: words don't have to denote objects. They can also denote concepts. (This is Greg Carlson's big idea.) If common nouns are the proper names of types, as proposed by Carlson (1977), and types are concepts, then language should be able to describe events like the activation of a concept in the mind of an agent, and keep such events distinct from events that take place in the world. The idea of experience mode is that it's a mode in which events happening *inside the brain* can be described.[6] So to sum

[6] Actually, **given that my whole idea is that sentence meanings are descriptions of SM sequences, everything in semantics describes episodes happening inside the brain. I think the main difference is that in internal perception mode we briefly turn off the interface between concepts and the world, so that SM sequences are guaranteed to describe events**

up: the proposal is that, in internal perception mode, when I say *John annoys me*, the SM routine that's described involves, at some point: (i) reference to the agent's cognitive representation of John (as opposed to the external individual John), and (ii) description of a purely mental event (the welling-up of a feeling) that this representation causes to happen.

11.2.0.4 Aside: real and simulated feelings

One problem for the above idea. Whan I say *The light dazzled me*, surely it's the actual light that dazzled me, rather than my cognitive representation of the light? And when I say *The knife hurt me*, it's the knife that does the hurting, not my mental representation of the knife? Maybe emotions are different, in that they can be as strongly associated with representations evoked in the absence of their physical denotata, while physical feelings like pain and dazzlement can only be intensely evoked by physical stimuli?

Obviously, thinking about a knife hurting me doesn't generate intense pain. But it does generate an emotion associated with pain: the concept 'knife' carries emotional connotations, and these do have something to do with pain, because they are learned when interaction with a real physical knife leads to pain. The emotion itself is not pain, but it's associated with pain. And I don't think it's just any old negative emotion: the emotion associated with a knife (which relates to slicing pain, or piercing pain) is different from that associated with a hammer (which relates to blunt-instrument pain) or a match (which relates to burning pain). I like the idea that these emotional connotations of objects are distinct, and mirror the actual feelings the objects can induce.

11.2.0.5 External perception mode and internal perception mode

I'll use the term **external perception mode** to talk about the mode where the observer's SM experience is yoked to the external world, and **internal perception mode** to talk about the mode where the observer has *disengaged* the interface linking his SM representations to the external world, and is 'observing' (if that's the right word) events that take place purely within his own brain.

I like the idea that the observer can observe events within his own brain.[7] This is entirely consistent with the idea that the meaning of a sentence is a rehearsed sequence of SM operations, with associated reafferent SM representations. When concrete episodes in the world are apprehended, the SM sequence obviously results from the observer engaging with the world, but if the observer disengages with the world, then the SM sequence could describe a purely mental event.

In fact, I often assume this is possible. For instance, when an observer retrieves an episode representation from episodic memory mode, he reactivates a SM sequence which

that are *purely* mental, rather than events in which mental events are correlated closely with things going on external to the agent.

[7] For this whole discussion, I need to make lots of references to recent syntactic models: in particular, Jackendoff (1990); Bouchard (1995); Landau (2009).

comes purely from his own brain. In this case, of course, the origin of the sequence was an experience in the actual world: a sequence experienced earlier was stored, and is now being retrieved. But I'm envisaging that internal experience mode is quite different from this. The event that's being experienced is an event going on *right now*: it just happens to be taking place purely within the observer's head. So I want to distinguish very clearly between internal experience mode and other cognitive modes like episodic memory mode. What I want to investigate is how internal experience mode works: how an observer gets into this mode, and what its purpose is.

The idea that there are two alternative modes, 'internal perception mode' and 'external perception mode', chimes well with Menon and Uddin's work on large-scale functional brain networks (see e.g. Menon and Uddin, 2010). I like the idea that there's an emotional saliency map which sums the amount of emotion being evoked and compares this with the total strength of affordance-based object representations in the current external scene, and decides which mode to go into. I suggested this decision process is implemented in the anterior insula; this accords well with Menon and Uddin's idea that the anterior insula is implicated in a 'salience network' that integrates external sensory representations with internal visceral representations. Except I don't think it's a matter of integration; I think it's first and foremost a matter of mode selection. Again this chimes with Menon and Uddin's idea, because they suggest the insula is involved in a mode-switching process of some kind.

11.3 Experiencer sentences: most recent version

There are interesting differences between these sentences.

(11.9) John loved the chocolate.

(11.10) The chocolate enraptured John.

(11.11) John tasted the chocolate.

(11.12) The chocolate tasted delicious. [To John]

I want to interpret all these sentences as they might occur in a close-perspective narrative, whose protagonist is John. (Not as they might occur e.g. as the answer to a question.)

11.3.1 *John loved the chocolate*

As a preamble, I propose that internal perception mode is a mode in which semantic stimuli compete to be selected as 'the object of John's consideration', with one stimulus eventually being selected. In external perception mode, objects in John's environment compete to be selected. In internal perception mode, John chooses to 'contemplate' some particular internal stimulus. For instance, he may choose to think about the person he's in love with (Mary): she dominates his thoughts. I suggest that stimuli compete in this mode in relation to the strength of their emotional associations. Mary dominates John's thoughts because she's associated with strong feelings. (The strong feelings could also be of hate, fear, happiness: objects which evoke strong feelings are the ones we 'dwell on'.)[8]

Now recall that internal perception mode competes with external perception mode: in this mode, John will attend to objects in his periphery, with a bias towards those that advance his plans, or those that are perceptually salient. I propose that there's a meta-level competition between internal and external perception modes. Think of the anterior insula as holding a representation of the emotional significance of all stimuli—probably with a bias towards recently-encountered stimuli. I suggest it's an *alternative saliency map*, that competes as a whole medium against the regular saliency map encoding external objects.

In the scenario I envisage, John first attends to himself as a regular agent, in external perception mode. Then his external saliency map and internal saliency maps compete. I like the idea that a map is selected not based on a winner, but based on the overall amount of activity in the map: some sort of summing operation. In our selected scenario, the winning map is the internal saliency map: the most important things going on right now are emotional rather than physical. So at this point, John enters internal perception mode.

In internal perception mode, I suggest the combined emotional representations jointly activate a distribution of activity over the set of concepts. The concept that wins is the chocolate: through this process, John establishes the chocolate as a stimulus to contemplate. (This operation is exactly analogous to what happens in external perception mode, where the winning salient location brings about an action of attention, overt or covert, to a point in space, and as a result, activation of an object stimulus.)

Having activated the chocolate as a stimulus to contemplate, John can now *select a feeling*. There will be several alternative feelings associated with the chocolate: once the chocolate is established as a stimulus, these can compete against one another, so that the dominant one wins—in this case 'love'. It's only at this point that John becomes *conscious* of the feeling 'love'.

While a single feeling is emerging as the winner, I suggest there is a concomitant subjective experience that's described in language with the verb *well up*. While this experience happens, I suggest John is also re-attending *to himself* as the one in whom the welling-up happens. This is Damasio's point. (Note, for Damasio the *change* in the self is absolutely crucial: it's the key thing.) The self is the medium in which conscious feelings well up, so to experience a feeling welling up is to activate a representation of onesself. Less mysteriously, one might say that by having his concept of self active while the feeling is welling up, John learns to represent himself as a medium in which *emotional events* happen.

Finally John re-attends to the chocolate, and learns an emotional association with the chocolate. Importantly, when the chocolate first *elicited* an emotional response from

[8] Before internal perception mode is selected, there are binding issues linking stimuli to feelings: lots of feelings are activated by lots of stimuli, and there's no way of knowing which feeling is activated by which stimulus. But that doesn't matter, because at this point the whole mode is competing with the alternative physical perception mode. Once internal mode is chosen, there will be a way to address the binding problems, that involves selective attention to a single stimulus and to a single feeling.

him, it needn't have been a learned one: it may be what behaviourists would call an 'unconditioned' response. John has to learn this association, so that the chocolate will thereafter be associated with a pleasurable feeling. This is learning a conditioned response to the chocolate.[9]

Why does John have to reattend first to himself, and only afterwards to the chocolate? Perhaps he attends to himself while the emotion is 'arriving', and attends to the chocolate when the emotion is fully established. In this scenario, John, the emotional agent, is the agent of the 'welling-up' of the feeling, while the feeling itself is associated with the chocolate. Obviously (as I've said before), there are two types of links connecting feelings with individuals: feelings have to be linked to experiencers, and to stimuli. These have to be kept completely separate. My suggestion is that this happens in the above way: the welling-up of the feeling is associated with the agent, and the feeling itself is associated with the stimulus.

Where, and how, in the system is the 'welling-up' of a feeling represented? I think this has to do with competition between feelings, once a target has been selected.

My idea is that a conscious, attended-to feeling is just *axiomatically* mine. Again, refer to Damasio. The medium of conscious feelings is the simplest medium in which we develop a concept of self. (I think Damasio calls it the 'core self': it's not extended in time, just experiential: even babies and animals are supposed to have it.) The main thing that happens in *John loved the chocolate* is that this core self, the feeling self, is *linked* to another representation of self: the agent of arbitrary actions, including motor actions, and the possessor of arbitrary properties, including both emotions and physical properties.[10]

On the above account, it's important that the 'consequent state' of the welling-up of an emotion is identified, since this functions as a transition point between John's attention being on himself-as-experiencer and on the chocolate-as-emotionally-significant-stimulus. So how is the consequent state of the welling-up cognitive episode identified? Well: the consequent state of a cup-grabbing action is when the agent's hand reaches the cup and the cup is re-established in a new modality (the haptic modality). Maybe the arrival of an emotion involves some settling, and the consequent state of its arrival is when 'a clear emotional signal is achieved'. That's my guess for now. I suggest this is the point at which an association can be learned linking the emotion to the chocolate-as-stimulus.

[9] In practice of course John's response to the chocolate is a mixture of conditioned and non-conditioned. But the presence of an unlearned response in this mixture is what allows John to learn new associations, or refined associations, towards the chocolate. It's the *unexpected* feeling that generates new learning.

[10] Importantly, this all goes for *observed*, third-party experiencers, as well as for onesself. Say the observer is me, rather than the one who's actually having the feeling (John). The general idea here is that I represent John as an experiencer of emotions at the point that I *empathise* with John. Specifically: I first attend to him as a regular external agent. Then I notice his generally heightened emotions and so enter internal perception mode (but now I'm 'perceiving', or rather inferring, *someone else's* internal brain states). Then I work out that John is activating *the chocolate* as a stimulus to contemplate. (Perhaps by joint attention, perhaps by inference, perhaps by language. E.g. he may be ecstatically saying "The chocolate, the chocolate!".) Then I parse the content of his emotion. My suggestion is that *during this process* I'm reattending to *John* as an experiencer of emotions. This is the empathy bit. Then I associate the feeling 'love' with the stimulus 'chocolate'—but *in a mental space associated with John.*

In summary: John attends to himself. In this state, two saliency maps compete. One relates to the physical world, and contains pre-attentional representations of objects (proto-objects). The other relates to the emotional world, and contains pre-attentional representations of feelings. (Collectively this can be called the proto-self, and the individual feelings can be called proto-feelings. The medium that holds these representations is the posterior insula. In this medium there's no competition between emotions.) These two media compete, based on their aggregate activity: in this case, the emotional saliency map wins. This means that emotional computations are of more importance than physical actions, and John moves into internal perception mode. In this mode, the first thing to happen is that an object for emotional analysis is selected. All concepts—call them stimuli—compete to be selected, regardless of their physical presence in John's situation. What's on his mind?! A single concept, that happens to be an object—the chocolate—is selected. Now that it's selected, a single feeling is selected, as the dominant emotion evoked by this selected object. This happens in a competitive medium—the anterior insula—where feelings compete against one another.[11] The process of a single feeling emerging as the focus of attention is called the 'welling-up' of a feeling. (It can be associated with a behaviour, like singing, which can likewise be said to well up. In humans, this behaviour once identified can be volitionally suppressed before it actually emerges.) While the feeling wells up, John activates a representation of himself as an agent in whom emotions can arise. When the feeling has finished welling up—i.e. is stably represented, for some finite period of time, before being inhibition-of-return happens—an association is learned between the feeling and the stimulus. The magnitude (and sign) of the change made to the association depends on the difference between the expected feeling and the actual feeling.

11.3.2 *The chocolate enraptured John*

I'll first give equivalent sentences in French and Italian, to show the object (*John*) differs from a normal object. The difference can be seen in the form the object pronoun takes. In Examples 11.13 and 11.14, the pronoun is glossed as *to-him*, while in the normal transitives in Examples 11.15 and 11.16, it's glossed as *him*.

(11.13) Le chocolat lui plaît.

(11.14) Il cioccolato gli piace.

(11.15) Jean le mange.

(11.16) Gianni lo mangia.

You can also use a PP, postposed or preposed, to convey the experiencer.

(11.17) Le chocolat plaît beaucoup à Jean.

[11] I think that the selected stimulus activates a bundle of feelings in the posterior insula, and these are projected one-to-one into the anterior insula, where a single one is selected.

(11.18) A Jean le chocolat plaît beaucoup.

(11.19) Il cioccolato piace molto a Gianni.

(11.20) A Gianni il cioccolato piace molto.

In fact you can present the PP in the middle as well:

(11.21) Le chocolat à Jean plaît beaucoup.

(11.22) Il cioccolato a Gianni piace molto.

In English we can't say 'To John chocolate pleases': but we can say 'To John, chocolate gives/engenders pleasure'. I think the Romance verbs *plaire* and *piacere* really function as intransitives. They are definitely causative as well: in fact they can be restated explicitly as causatives (*plaire* is the same as *faire plaisir*; *piacere* is the same as *fare piacere*).[12]

So the English verb *enrapture* means 'give rapture to', or 'cause rapture in'. The English verb includes the preposition 'to' in its meaning, while this is conveyed in the object in the Romance constructions rather than the verb.

I think I want *The chocolate enraptured John* to mean 'In John, the chocolate engendered rapture'.[13] In this case we don't have the structure of a transitive SM routine at all. The structure involves something more like a locative PP, which can pop up in arbitrary places in the sentence (and does in Romance). My first guess is that there are two SM routines that run kind of in parallel. In one, John's mental world is established as the environment for the episode currently being monitored. (I think this conveys John's adoption of internal perception mode: something that in *John loved the chocolate* is conveyed implicitly, in the AgrOP projection.) In the other, chocolate engenders rapture. But I think there's *some* important relationship between these two routines.

For one thing, once the emotional event is construed as happening 'in John', feelings become *objects* rather than actions. The only action in *The chocolate enraptured John* is the action of engendering (causing the welling-up of) a feeling. In *John loved the chocolate*, the content of the feeling is conveyed by the open-class main verb. The decision to situate the action in John totally changes the way emotional events are described: specifically, it changes the linguistic system that reports the emotions. Understanding how this happens is key to distinguishing between *John loved the chocolate* and *The chocolate enraptured John*.

The above reasoning sugggests that selecting John's mental world as the 'current environment' happens *first*. But if that's the case, then two questions arise: first, why isn't John the subject of *Chocolate enraptured John*? And second, why can the PP *in John*

[12] In English there's often an explicit causative reformulation of experience verbs as well: for instance *The fire scared John*, *The fire made John scared*.

[13] Note we can also say *The chocolate engendered in John rapture* or *The chocolate engendered rapture in John*. The middle PP is a bit weird unless the emotion noun is heavy—for instance *engendered in John a feeling of rapture*.

show up not only preposed, but also mid-posed and postposed in the sentence *Chocolate engendered rapture*?

One way to approach that is to think about how PP modifiers work in general. In *John gave a cup to Bill* the PP can also appear in several places (*To Bill John gave a cup*, *John gave to Bill a cup*). But the trajectory *to Bill* features at a single point in the SM sequence being described, I think. (At the end, somewhere.) My idea in this case is that the PP describes a control system that's kind of autonomous or self-contained, in that it has its own coordinate system: in this case centred on the cup's environment. (In *John walked to the door*, centred on John's environment, represented allocentrically.) The idea is that the controller operating in this allocentric environment controls the *parameters* of a motor controller operating in a motor environment—for instance, the goal locations to be reached by the hands and feet during walking, which are different e.g. when going straight and turning a corner. Thus the PP describes a *meta-level* controller.[14] This suggests that the PP plan can be stored as a separate plan, and there are several opportunities for its replay when the clause-level SM sequence is being replayed, in much the same way that a DP is stored as a prepared plan in a separate medium, for which there are multiple opportunities for replay. The benefactive PP issue is then: why can you read out the full PP at the start of the clause, and before the object, and at the end of the clause? Turning to the experiencer PP *in John*, the analogous proposal would be that this PP describes a meta-level control operation, that happens in its own coordinate system. I don't think this idea gets us anywhere.

It's interesting that the pronominal *in him* shows up as a clitic in Romance. (*Le chocolat lui plaît.*) In my model, this means the establishment of John's mental world as the current environment can somehow be read out from the planned SM sequence stored in working memory. Does this mean there's some particular time when this establishment happens? If so, why can it be expressed in three places in the sentence when it's a full PP? [15]

Think about the causative structure for a moment. It's plausible that *en-* and *rapture* are two heads, denoting a causative operation and activation of a feeling respectively. But it's far from a normal causative structure: *rapture* is not a verb, and the thing doing the causing is a stimulus.

'The chocolate' (either as object or stimulus—I think probably stimulus) has to be perceived as the agent or instigator of a causative process, which results in a feeling welling up in John. The only action it can do is to elicit, or engender, a feeling in an experiencer. How can a stimulus be registered as doing this?

One proposal is as follows: to say a stimulus causes a feeling to well up is to say that in the past, activation of that stimulus has reliably led to welling-up of that feeling, so when the stimulus occurs now and the feeling wells up shortly after, we assume that the stimulus is *responsible*. After all, it's important that the brain identifies the stimulus that *actually*

[14] More on this in Section 11.4.

[15] It's also interesting that the pronoun is the same as for benefactives (e.g. *Jean lui donna la tasse*). I argued in Section 10.1.1 that benefactive arguments denote experiencers; maybe there's some link here. Causative volitional actions like 'give her a cup' can have an impact on an experiencer; so too can the 'actions' of stimuli on an experiencer experiencing feelings in internal perception mode.

led to the feeling that welled up, and strengthens associations between this stimulus *in particular* and the feeling that welled up.[16] If we only have one stimulus active, and see what feeling wells up, this helps work out the causal role of the stimulus. Don't forget: linking a stimulus to a feeling is a complex piece of conditioning, that behavourists know heaps about. The general idea I'm proposing is that a central notion of causality features in the architecture of the network that implements instrumental conditioning (that's the work with Jeremy) and also in the network that implements classical conditioning (that's the learning of conditioned responses to stimuli).

In my model of causative actions with Jeremy, the concept 'cause' denotes the entering of a special mode, implemented by a special network, in which motor actions are associated through learning with arbitrary perceptual effects. Is there a similar operation in the scenario where a stimulus causes a feeling to well up? Well, a stimulus doesn't always cause a feeling to well up. This only happens if the stimulus is attended to in internal perception mode. In this mode, as discussed in the section on 'John loved the chocolate', (i) a single stimulus is selected, and (ii) competition is instigated between alternative feelings (in the anterior insula), so that a single one 'wells up'. Possibly the causal operation in the meaning of *enrapture* (analysed as =*cause-rapture-in*) denotes *activation of the circuitry that initiates competition between feelings in the anterior insula*. But has that got anything to do with the concept of cause as used in causative motor actions? There's the notion of arbitrary learned associations—but that seems rather weak/minimal as the basis for a neural account of the concept of causation.

A useful survey of current models of the concept of causality as it applies in models of conditioning is given in Waldmann *et al.* (2008). Identifying that X causes Y involves a lot more than just recognising a constant association between X and Y: minimally X has to come *before* Y (and lots of other conditions too).[17]

Maybe I'm focussing too much on the stimulus, rather than the physical object. Don't forget that when we first perceive a stimulus, we don't know all the associated feelings: we have to learn these associations. For instance, a child doesn't know the association between the category 'chocolate' and the feeling 'gustatory-pleasure'. So the causal event should almost certainly be the stimulus *at the time it's actually presented to the experiencer*, rather than at some later time when it's presented in isolation.[18] At the time of actual presentation, John activates the chocolate concept, and also the feeling—but these two aren't yet linked. The idea is that *The chocolate enraptured John* describes a learning episode, where the observer strengthens an association between the chocolate and the feeling 'rapture' (within the environment of John's mental processes). (The idea is: he

[16] In very low-level terms, STDP makes sure that a connection is only strengthened if the presynaptic spike occurs *just before* the postsynaptic spike.

[17] Another factor is *time*. When we see a stimulus that predicts a reward in the future, we get a dopamine burst a specific time into the future. But I don't think that when we say *The book delighted John*, we're describing a process that takes some arbitrary amount of time.

[18] Now I'm thinking about *The chocolate enraptured John*, not the generic *Chocolate enraptured John*. I guess the generic summarises the results of learning, and perhaps does some consolidation, but doesn't do any *new* learning.

only strengthens this association because he has some evidence that the chocolate stimulus *causes* the feeling: minimally, the feeling happens afterwards, after an appropriate delay.)

So much for 'the chocolate caused rapture to well up'. Now return to the relation between this causal process and the establishment of John's mental processes as the current environment (*in John*). What can I learn from the fact that in English, the verb *enrapture* (or equally *intrigue*, *horrify* etc) includes the 'in' predicate, whereas in Romance, the verb (e.g. *plaire*) doesn't include this predicate? Because you say *Y plaît <u>à</u> quelqu'un.* It's obviously possible to divide things up in two ways.

One possibility is that *John* means 'John *the experiencer*'. For instance, it could signal the kind of attention to John that's associated with entering internal perception mode. Those two things together would indicate that the observer is now

Think about how an observer external to John would perceive *The chocolate enraptured John.* The observer first attends to John: I'm pretty sure. Then he identifies John as an experiencer, because he recognises that John is 'having feelings' rather than doing something. (He doesn't recognise any *specific* feelings at this point: all he does is recognise that he should go into internal perception mode, so he can perceive (or rather infer) the feelings of the currently-attended external agent.) I'll call this mode **empathy mode**.[19] All the above is the process denoted by 'to John', I think. (This is the bit that's implicitly represented in *the chocolate* in *John loved the chocolate.*)

The rest of the SM process is what happens in empathy mode. Empathy mode is implicitly referred to an external agent: the agent has already been identified, and the observer has worked out that empathy mode needs to be engaged: the only thing remaining is to work out (a) what the observed agent's emotion is *about*, and (b) what the emotion *is.* I suggest the observer in empathy mode first identifies what the observed agent's emotion is *about.* (The observer already knows that the observed agent is experiencing *some* emotion, so it makes sense that the emotion-classification mechanism gets some information about the likely cause of the emotion before coming to a conclusion.)

I suggest that finding out what the cause of the observed emotion is involves a *special form of attention-following.* John can follow Mary's visual attention, as normal, because Mary might be looking at the thing that's generating her feeling. But Mary's feeling might also be generated through touch, or taste, or audition: I suggest that John also looks at what Mary is touching/feeling, or at what she is eating/smelling, or what she is listening to, to come up with a hypothesis. I suggest John can also listen to the content words spoken by Mary. (Perhaps this process is different from normal sentence interpretation. It might involve listening mainly to the interjections, or words spoken with emotion: *Oh this <u>chocolate!</u>*) The upshot of all this is that John identifies a single stimulus that Mary is attending to, that is likely to be the cause of her emotion. Again what's happening here is that stimuli are competing, not through physical salience—or at least not *just* through physical salience—but also on the strength of their emotional associations. Things with

[19] This is quite subtle. I'll use the term **feeling mode** to refer to internal perception of one's own feelings, and **empathy mode** to refer to internal perception of someone else's feelings. So we still have a distinction between attend-to-self mode and attend-to-other mode, and in either mode we can opt to enter internal perception mode.

high emotional associations have higher 'emotional salience' in the 'emotional saliency map'.[20]

Finally, having identified the item that Mary's emotion is likely to be 'about', the observer's emotion-recognition system classifies the emotion itself. Now it's identifying one single emotion. In our example, the identified feeling is rapture. The emotion classifier takes the observed agent's face and body-language as input, obviously, but also her utterances (including semantic content, but also nonlinguistic speech things like laughs, cries etc). My idea is that it also takes the item the feeling is hypothesised to be 'about' as input: the important thing is that the emotion is plausibly something that's brought about by this item. (So the causative/engendering process is centre-stage in the mechanism that identifies feelings in other agents, just as it is in the mechanism that identifies feelings in onesself.) One mechanism is that the hypothesised 'object' of the emotion creates a distribution of possible emotions (e.g. based on the observer's own experience), which provides top-down input to the emotion recogniser.

The emotion-recognition mechanism outlined above is nice, in that the sequential structure of emotion-recognition is the same for one's own emotions and for those of an observed external agent. This idea mirrors what happens in my model of action execution and action perception. It's also a nice account of empathy. What the observer is doing is having a feeling *in response to an identified object*—or at least simulating having this feeling.

11.3.3 *The chocolate enraptured John* again

The observer first attends to John. (John is either himself or an external agent.) Now three neural systems compete: (1) a system for identifying the currently-attended agent's feelings; (2) a system for monitoring the currently-attended agent's actions; (3) a system for identifying the currently-attended individual's static properties. In this case, the first system wins, because the observer has high overall activity in his emotional saliency map. So the observer enters feeling perception mode. This mode is deictically referred to the attended agent, who is either the observer himself, or an external agent.

In feeling perception mode, the observer first identifies what I'll call the **object** of the feeling: a process involving competition in an emotional saliency map. (In attend-to-self mode, each candidate object is activated in measure of the strength of its emotional associations, and the winner is attended to. In attend-to-other mode, emotional associations of objects create a top-down bias, but the winner is selected by identifying the observed agent's actual focus of attention.) The observer then identifies the feeling associated with this object. (In attend-to-self mode, this involves selection of the winning feeling, plus some kind of *check* to make sure that the object *caused* this feeling. In attend-to-other mode, identifying the feeling involves classifying the observed agent's expression, body language,

[20] Recall: I already postulated an emotional saliency map in the anterior insula for an agent's *own* feelings. Now I have to find some way of recruiting this same map to support the identification of the stimulus that's most emotionally significant for an *observed agent*. In this role, the emotional saliency map probably has a spatial component, like the physical saliency map.

utterances, and again the *check* to make sure that the object can plausibly be understood as the cause of the hypothesised feeling.)

Two questions about feeling-perception mode. First, what's the relationship between the process of attending to the feeling and the 'check for causation' just described? Secondly, is the 'welling-up' of the feeling perceived in the feeling-perception process, and if so, how does it participate in the process? As for the second question, I think that welling-up is experienced during the time when the feeling is emerging as the winning feeling. (That's what I said in my account of *John loved the chocolate*, anyway.) As for the first question, I can think of two possibilities. One is that straight after identifying the object of the feeling (the chocolate), the observer executes 'cause' and then activates 'rapture'. More specifically, after identifying the chocolate, the observer executes a control operation that engages a network encoding cause-effect relationships, resulting in activation of a distribution of feelings that could potentially be caused by this object. (This network-enabling operation could be the denotation of the word *cause*, or maybe the clitic *en-*.) It's only after this is done that a particular feeling is activated. The other possibility is that after the observer attends to the chocolate as a stimulus, he first identifies the dominant feeling (rapture) and only afterwards works out whether the feeling is plausibly caused by the chocolate. I don't think the latter option is right. In a causative motor action, the agent attends to the object, but then brings about an effect on this object: in the current case, the chocolate isn't bringing about an effect *on* the emotion as an object, it's causing the *appearance* of the emotion (its welling-up). This makes me think that the former option must be right: the chocolate doesn't provide input to the causal process, rather it's the *output* of this process. Note the situation is very different from volitional verbs of creation: there it's plausible that a motor programme representing (e.g. a square) is the input to the creation process, so it's not created from scratch. (It couldn't be.) So this is a very different kind of causation. We're literally interested in what feeling the stimulus causes to appear. There's no notion of volitionality.[21]

11.3.3.1 SM interpretation of the syntax of *The chocolate enraptured John*

The syntax of the sentence has to convey two linked SM processes. The first is the process of attending to John (the physical agent), identifying that he's experiencing feelings, and entering feeling-perception mode. The second is a process, deictically referred to John, of identifying (i) the object of the feeling (i.e. what it's 'about') and (ii) the category of the feeling (e.g. rapture), and of checking that the object plausibly causes this feeling. I'll call the first process the **feeling-mode-establishing process**, and the second process the

[21] Note: in English the causative suffix *-en* can be applied to nouns denoting feelings (e.g. *fright-en*='cause fright in') but not to verbs (e.g. *learn-en* doesn't mean 'cause-to-learn'). Maybe there are two types of causality, and English *-en* only signals one of them. However *-en* can also be applied to quite productively to adjectives, with the meaning 'cause-to-become-more-X' (e.g. *quicken*='cause-to-become-more-quick'); I'm not sure what this category of causation has in common with the feeling-causing kind, but *quicken* can be used as a causative verb, so This question relates to Pylkkänen's discussion of the parameter determining what CAUSE applies to; see Section 13.0.6.

feeling-identification process. So we have to look for syntactic structures conveying these two processes, and syntactic structures conveying the relation *between* these two processes.

In the French *Le chocolat plaît à Jean*, I suggest the feeling-mode-establishing process is expressed in the PP *à Jean*, and the feeling-identification process is conveyed in the intransitive clause *Le chocolat plaît*.

How does the PP *à Jean* signal the feeling-mode-establishing process? My suggestion is that the P *à* describes the SM operation 'attend-to-Jean-as-environment', the specifier of PP describes the reafferent stimulus 'Jean-as-environment', and the complement of PP describes the state the observer gets into after having attended to Jean as an environment (i.e. the state in which Jean is the active environment, and we're in feeling-perception mode).[22] I like the idea that the preposition *à* signals an action of attention to an agent *as an environment*, or more accurately *as an experiencer*. When I walk *to the door*, I attend to the door as a place—but in doing so I also attend to a particular *location*, and this allows the specifier of the PP to function as an index to a regular DP, just like case-assigning positions in the clause. So a P is an action of attention to a thing, just like AgrS or AgrO. And just as for AgrS/AgrO, I assume that the planned action of attention includes reference to the thing, but this can't be expressed in language: all that can be expressed is certain properties of the planned action of attention itself.

Syntactically, the PP is interestingly decoupled from the main clause: it can appear preposed, embedded or postposed in the clause. I have two suggestions about the nature of the relationship between the two SM processes, and how it's conveyed syntactically. Firstly, I propose that the two SM processes are *stored separately in WM*. Secondly, I suggest that even though the PP contributes the primary mode-setting SM process, in *Le chocolat plaît à Jean* the *rehearsal* process begins with the feeling-identification process (the one denoted by the main clause). During the *structuring* rehearsal process, an opportunity must be found to replay the feeling-mode-establishing process as an embedded sub-process. There are several opportunities that can be taken. This would explain why the PP doesn't syntactically dominate the structure of the clause.

This is a really interesting general idea: during a SM experience, several somewhat distinct SM routines can be executed, and stored separately in WM. During replay (e.g. for sentence generation), one of these is the **structuring SM routine**: this means that while this routine is being rehearsed, opportunities must be found to rehearse the other SM routines as 'embedded' routines.[23]

Why would rehearsal be structured around one SM process rather than the other? It

[22] Why does this whole PP need to be separate from the main clause? Well, it's the same in physical PPs. Consider *In France, they eat bavette aux échalottes*: the PP identifies where the proposition is experienced; the clause identifies the proposition. We can just accept that the location of the experience is delivered by a separate SM process.

[23] Note: this idea provides a nice general framework within which we can give an account of how *DPs* are embedded within clauses. In particular, recall that a DP pronoun can be realised within the main clause as a clitic—so can a pronominal PP in Romance (e.g. the French *lui*, *y* etc). This could be a helpful way of thinking about *Le chocolat lui plaît*.

could be a matter of convention in a given language. But it might also be that there's flexibility within a language. This is something I'll have to think about.

Why are there exactly three opportunities to rehearse the feeling-mode-establishment process within the feeling-identification process? I'll take my lead from the idea that DPs denote attentional SM routines that are *indexed* to points in a proposition-establishing SM routine. The link in this case is claimed to be through *cross-modal object representations.* Maybe it's the same for feeling sentences. Consider the structuring SM routine: 'Chocolate causes-to-well-up rapture'. (The structure of that is: attend-to-chocolate, activate-cause, experience-well-up (with experiencer as reafferent side-effect?), attend-to-rapture.) If the operations activate-cause and experience-well-up each generate are representation of the experiencer (Jean) as a reafferent side-effect, one possibility is that cross-modal associations between Jean the individual and Jean the environment allow the PP *à Jean* to appear in these specifier positions. Maybe what allows it to appear at the start is that the *state* denoted by AgrSP of the main clause is one that indexes the PP. Maybe modification at the end of the clause happens the other way round: the PP *à Jean* is the dominant structure, and the whole clause is read out as an optional modifier of this. (I'm thinking that the clause could somehow be indexed to the consequent state of the PP, so that rehearsal of the PP provides an opportunity to read out the clause.)

In the English *The chocolate enraptured John* it's more complicated: it's partly expressed in the surface object *John*, but also in the *en-* morpheme of *enrapture.* I can't figure out why it's okay to decompose things this way.

One idea is that in English, the feeling-mode-establishing process is entirely unsignalled, with the DP *John* being read from the specifier of the VP describing the welling-up of the feeling. Like Damasio, I'm proposing that while a feeling is actively welling up, a representation of the experiencer 'as an experiencer' is activated: through cross-modal links, this must be connected to a representation of the experiencer as a physical individual. This means that the specifier of the VP headed by *well-up* is an opportunity to rehearse the DP describing the experiencer 'as an object'.

This is an interesting idea, but it presupposes quite a lot about the parameters of variation for expressing SM processes. In the general case, reporting a SM experience involves conveying a set of SM routines. I'll ignore the DP-denoting routines for now: in our example there are two SM routines to convey. To be communicatively successful, a linguistic convention must allow both these routines to be conveyed. One way of doing this is the French way, where the feeling-mode-establishing routine is conveyed explicitly, and indexed to the feeling-identification routine. Another way of doing it is the English way, where the feeling-mode-establishing routine is not conveyed explicitly at all, but by indexing a *DP* denoting the experiencer to the feeling-identification routine, the speaker does enough to allow the hearer to *recover* the full feeling-mode-establishing routine. An account of how this happens obviously requires an account of sentence parsing, which I haven't yet considered. But I'm already committed to the idea that parsing is very different from generation. Parsing involves (i) hypothesising a set of SM routines based on surface word-sequences within the sentence; (ii) working out how they can be interrelated. I'm going to assume it's possible to recover a fully implicit SM sequence in a case like the one above, and leave the question of how this happens for future work.

11.3.4 *John tasted the chocolate*

This may seem like an ordinary transitive action sentence: but *taste* is a **perception verb**, so it's not quite ordinary. To analyse this sentence, we need to set it alongside sentences conveying other perception verbs.

(11.23) John saw the chocolate.

(11.24) John heard the chocolate.

(11.25) John saw the chocolate spill.

(11.26) John heard the chocolate spill.

(11.27) John saw [that the chocolate was hot].

(11.28) John heard [that the chocolate was hot].

It's hard to taste events, but just about possible in some cases:

(11.29) John tasted the wine oxidise.

However, tasting is a means by which we can perceive propositions (about the tasted object):

(11.30) John tasted [that the chocolate had salt in it].

This means that tasting has to be analysed in the same way as the more productive perception verbs like seeing and hearing.

I already have a suggestion about how these perception verbs work (see the 'Verbs of perception' section in Part 2). My suggestion is that John first establishes himself as the agent, and then attends to the chocolate—and then, in a nonstandard or 'interrupting' SM operation, establishes perception mode. When perception mode is established after John is established as the agent, and has attended to an external object, the mode-setting operation is reported *as a verb*.[24] The nature of the verb depends on the modality in which perception happens. If the agent percieves the object primarily through vision, the verb is *see*; if, as in the current case, the agent perceieves the object primarily through gustation, the verb is *taste*.

I suggest that as a side-effect of the operation of entering (gustatory) perception mode, the agent re-attends to himself as an agent in a new modality. What is this modality? I don't think the agent is represented as someone in whom a particular feeling can 'well up': tasting is an action, not a feeling. (The content of the feeling generated by tasting is

[24] When it's established 'in the initial context', the operation of establishing perception mode is reported in AgrSP, and is conveyed in surface language in the agreement features of the subject.

represented by the adjective *delicious* in this case.) I think the modality is just the agent as an animate object—an individual that can undergo certain patterns of movement or change—rather like the animate agent representation evoked when regular motor actions like running and grasping are under way. The main difference is that here the change is in the *mode* of the agent—indicating that the agent is the kind of being that can change its mode, maybe. For more detail on this side-effect, see the 'Re-attention to self' paragraph in Section 13.0.2.

At the end of the SM routine, the chocolate is re-attended to 'as a feeling': this is the point when the affective connotation 'gustatory pleasure' is associated with the representation of the physical chocolate, so that in future it has (some of) this sensory/affective association. That's all quite straightforward.

11.3.5 *The chocolate tasted delicious*

Here the feeling-mode-establishing process is again clearly implicit—unless the PP *to John* is added. (Note it can be added at several places, as in French.) The clause reports the feeling-identification routine. But here I suggest it reports this routine as a process of attending to the *properties* of the chocolate, rather than a process of attending to the feelings that the chocolate causes to well up. The chocolate is a big gestalt of properties: importantly, alongside visual properties there are emotional properties. The sentence *The chocolate was delicious* reports on the process of identifying the property *delicious* as an idiosyncratic property of this piece of chocolate: it's more delicious than normal, prototypical chocolate.

The process through which this is established involves (i) categorising the attended object (as chocolate), (ii) inhibiting the identified category, and (iii) seeing what properties (if any) are left. In this case, the property that's left is deliciousness.

There are two things to explain. Firstly, why is the main verb *taste* rather than *be*?[25] Secondly, why is the content of the feeling reported as an adjective, rather than as a verb (like *love* or *enrapture*)?

I think the answer to both questions turns on the idea that the sentence reports a process within the property-recognising system, which is part of the **object classification system** rather than the episode apprehension system. I suggest that when an agent experiences an emotion, SM routines are detected within the object-classification system *and* within the episode-apprehension system; these two routines correlate quite strongly, so when the experience is conveyed in language, they don't both need to be reported in every detail.[26]

As to the first question. I suggest that when John identifies a certain feeling as an idiosyncratic property of the chocolate, he *attends to the feeling*, in exactly the same way as the feeling is attended to when it's reported in the causative verb *en-rapture* And the process of attending

As to the second question. I suggest that the content of the feeling is reported as a

[25] It certainly *can* be reported with *be* as well: no problem there at all. And if it is, the hearer will easily infer that identifying the deliciousness involves a process of tasting it.

[26] The story is just the same as for the distinction between English *S pleased E* French *S plaît à E.*

11.3.5.0.1 *To-John [something] caused-to-laugh*

11.4 Experiencer PPs

In this section I want to consider in more detail the role of the PP *à Jean* (or its English equivalents *to John/in John*) in *Le chocolat plaît à Jean.* The aspect I want to focus on is that the phrase signals an operation that the material in the main clause is deictically referred to. My feeling is that this is even the case in a sentence like the following, where the PP seems to be embedded within a VP:

(11.31) John walked *to the door.*

In this example, I've always thought that the PP indicates that control of John's action is taken over by a meta-level controller working in an environment-centred coordinate system, whose meta-level 'actions' are changes to the parameters of the first-order motor action 'walk' (that's defined in a body-centered coordinate system).[27] In this sense, the PP controller is deictically 'above' the VP—in kind of the same way that the PP *à Jean* is deictically above *Le chocolat plaît* (and has sentence-level scope).

11.4.0.1 Static and dynamic PPs

I want to think of a way whereby a PP can always denote the same SM operation, but where the operation has very different effects depending on when it's executed. Specifically, if it's done while a motor action is under way, it will result in something moving into a new location—something actually undergoing the described trajectory—while if it's done beforehand, the change is just an attentional one (e.g. attending to a particular place or situation, or in the case of *à Jean*, attending 'from the inside' to the inner workings of someone's mental processes).

11.4.0.2 Updates to 'the current environment'

I'm trying to think about the operation denoted by *à Jean*, which establishes Jean as a new environment (in which 'objects' like stimuli and feelings are situated, and in which 'episodes' can occur involving these objects).

One thing to ponder is that a PP always signals a *change* in the current environment. For instance, in *John sat on the chair*, we have to imagine that in the initial context, there's an active environment, in which John is an object. By the end of the sentence, the environment is the chair, and John is an object in *that* environment. (He's been *re-indexed.*)

When I say *John walked <u>to the door</u>*, the locomotion verb *walk* (or more precisely, the implicit *go* in this position) indicates that John's *environment* is being selected as the

[27] It's interesting to think of the PP as signalling an 'interrupting' mode-changing operation, somewhat like the one in *John saw the cup.* I'm pretty sure the PP signals a mode-changing operation—and since it occurs mid-way through a SM routine, I think it would count as interrupting. I think it *takes over*—but at the same time, it works *within the SM context active at the time it takes over.*

medium in which alternative motor actions will compete—rather than the 'default' environment, which is, I presume, John's peripersonal space (an environment defined relative to John's motor system). Having attended to John as an object in the initial environment (say the room), I guess the environment changes to a motor environment centred on John, and the natural thing to happen in this context is selection of a single goal motor state (which is denoted by AgrOP). But if we decide to locomote, John's allocentric environment wins instead. At this point we can also decide what body-level motor action to do ('walking'). But we still haven't decided what trajectory to follow yet. That's described in the PP *to the door*. I think the main idea here is that an action of attention to an object looks very different when it's controlled by an environment-centred location. I suggest the thing that generates an action of focal attention to a location is *selection of a trajectory*—which *includes* selection of a landmark—see Part 2—so that's the selection operation that's conveyed by the P head.[28] (This makes the attentional action signalled by PP very different from those signalled by AgrSP and AgrOP, which are *pure* attentional actions. In the PP case, I think you choose the object and the trajectory at the same time, because the object in some sense determines the trajectory.)

The point is: we're already committed to a locomotion action ('walking') when we decide on a trajectory: so the selection operation denoted by the PP actually influences the course of an action. In a different way to the way the operation denoted by AgrOP influences an action—but with a similar effect. A PP that establishes an environment before an episode is monitored won't have that dynamic role.

What I want to think about is: consider *Le chocolat plaît à Jean*. If *to the door* indicates selection of a specific trajectory in an environment (that defines a set of competing candidate trajectories), *what's the environment in which Jean is selected?* Does *that* environment have to be nonstandardly established? I thought that *Jean* was the environment. There seems to be a disanalogy between *to the door* and *à Jean*: in the former case, the door is a selected place *within* an environment; in the latter, Jean is *the whole environment*.

So: how is Jean selected as an environment? Are there alternatives? I think maybe I have to attend to Jean as an object *first*. Then my choices are: (i) See what physical episode happens involving Jean; or (ii) establish Jean as an environment. That's certainly my idea for *Jean loved chocolate*. In that case, we attend to Jean, then (implicitly) establish Jean as an environment, then attend to an object inside that environment (chocolate), then activate the feeling 'love', then reattend to Jean (as active experiencer of feelings), then reattend to chocolate (to associate the feeling 'love' to the stimulus 'chocolate'). But what about *Le chocolat plaît à Jean*?

In this case, my attention must be drawn to Jean as an environment somehow. How about if I'm Jean myself? I must attend to myself as an environment in which episodes involving stimuli and feelings take place. How is my attention drawn to this environment? Is it at the point when feeling mode wins? If so, how do I know the mode is representing *my own* feelings? This is a key question. One possibility is that I have already attended to

[28] A pronominal PP (*there*) can be cliticised to the verb—so can *to him/her* (*Le chocolat lui plaît*)—these facts have to be taken into account.

myself *as an object*—for instance, I have established action execution mode. If that's the case, then this action of attention *is not explicitly reported in the syntax* (of *Le chocolat plaît à Jean*).

Now think about the case where Jean is an external agent being observed. Here I can't see any way round establishing Jean first as an object, and then attending to him as a mental environment. Again, this would mean that there's an implicit, early, AgrS.

What happens to this early operation? Why isn't it seen in the syntax of *Le chocolat plaît à Jean*? One possibility is that the operation is *fully inhibited* when feeling mode is established, to make way for the episode experienced *within* feeling mode (reported by *Le chocolat plaît*).

The next question is: why is the operation establishing feeling mode reported as a PP, rather than as part of the episode experienced in feeling mode? The suggestion is that it's represented *in a separate WM system*, and therefore has to be integrated within the sentence structure via an *indexing* operation (a piece of indirection) linked to a signal in the SM routine.[29]

The key idea in that last paragraph is that the environment-establishing routine is stored *in its own WM system*: it's separately stored, and is separately replayable. But it's *indexed* in various ways to the episode-perceiving routine.

I think the idea that the environment-establishing routine is stored in its own WM system is *related* to the idea that PPs signal 'interrupting' or 'nonstandard' SM operations. Think about it: I attend to an agent (say myself). This sets up my motor saliency map as the active map, and competition starts to happen within this map. At the same time, it sets up my *emotional* saliency map as a competitor to the motor saliency map. If this latter map wins, something non-standard will happen: I'll establish feeling perception mode. I envisage two situations. In one, I don't inhibit the full WM episode: I continue to attend to myself as an agent, leave the mode-changing operation implicit, attend to chocolate-as-stimulus, then activate the feeling *love* (as a 'motor programme'), reattend to myself, and then to the chocolate.[30] In the other situation, I do inhibit the full WM episode, which somehow allows me to represent my own feelings as *objects* (with extra precision) rather than as actions. In this case I still need a way to *communicate* which person these feeling-episodes are happening in. In English stimulus-experiencer verbs this is done by indexing a stored action of atttention to the experiencer. In Romance emotion verbs this is done by indexing to the separately stored

Note, in this latter situation, I do something non-standard: I deviate from the default path. I suggest this means executing an operation in a SM system separate from the episode-perception system, that's recorded in its own WM medium. Even then there's still a choice, about whether to start from scratch in the episode-perception system (yielding

[29] Which could be a transiently active one, in which case we get a PP or DP, or it could be a tonically active one, in the special pronominal cases where the tonic signal carries enough information. Or perhaps both, if there's 'clitic doubling'.

[30] Somehow I have to explain why the system of motor programmes only has coarse-grained access to feelings: the only verbs we can use are *like*, *love*, *hate*, which give us valence and strength but no actual content.

The chocolate caused pleasure), or to continue (yielding *John loved the chocolate*). The latter case results in what I call an 'unreported' mode-changing operation. We attend to John, *then we change to experience mode*, then we attend to the chocolate.

11.5 Nonstandard SM operations and their syntactic reflexes

Thinking in general about these nonstandard SM operations, I can think of various ways they happen, and various ways they surface in language.

1. Some mode-changing operations occur as motor operations, part-way through a SM routine, and are reported as verbs. E.g. *say* ('enter verbal mode'); *see* ('enter external perception mode'). Perhaps *feel* (as in 'feel sad'—'enter feeling mode').

2. Some mode-changing operations happen externally to a SM routine, some way through it, but not in a position where they can be reported (e.g. *John loved the chocolate*). Here we enter feeling mode after establishing John, and before establishing the chocolate (as internal stimulus). In these cases, the mode-changing operation has an impact on the remainder of the routine—like case 1: the routine reports things that happen partly in one mode and partly in another—but unlike case 1, the operation itself isn't reported in the sentence, but is left implicit. (Recoverable from the initial establishment of John.)

3. Some mode-changing operations happen some way through a SM routine, and for some reason the interrupted routine is fully *inhibited*, so that the episode that's then experienced occurs *fully* within the new mode. This is the case for *The chocolate delighted John*, and *The chocolate caused delight to John.* In these cases, there's a choice about whether you report the feeling as a verb or noun, and concomitantly, a choice about whether you include reference to the mode-changing operation as a DP or a PP. But in either case, the mode-changing operation is reported as a kind of adjunct, I think.

Maybe these same options also help explain what's going on when you decide to do a locomotion action. Having selected yourself as an object *within* the environment (the room), and created a motor environment centred on yourself, the normal thing to do would be to do an intransitive action (straight away), or a transitive action (by selecting an item within the newly-current motor environment). But instead you decide to re-establish your environment (the room) as a locomotion environment. The environment competes *as an environment* against the already-selected motor environment, so if it wins, it does so non-standardly. Having said that, we don't inhibit the SM sequence executed so far—in fact we continue: we select an intransitive motor action ('walk'). We then select a trajectory in the environment-based SM system: 'to the door'. Selecting this trajectory has a special kind of impact on the walking action: basically it 'steers' the walking action. While this is happening, there are two SM routines going on in parallel. One's reported by the VP headed by *walk*; the other by the PP *to the door.* This example somewhat resembles case (2) above, in that at the time when the room environment is nonstandardly selected as the new environment, the SM operation that's reported is the selection (in the default motor environment) of the intransitive motor programme ‘walk’.

What *generalisation* can I make about PPs? Syntactically, PPs are a Case-assigning mechanism: a P head licenses a DP.

11.5.1 *About* and *with*

The prepositions *about* and *with* are interesting. Consider these sentences, for example:

(11.32) John was cross about Mary.

(11.33) John was cross with Mary.

If John is cross *about* Mary, that means she *makes* him cross. If John is cross *with* Mary, he volitionally *expresses* his crossness *to* Mary. (He can do both these things at once: he can be cross *with* Mary *about* something she did.)

In each case we have an adjunct PP: a PP that provides additional information. And this concerns the establishment of another environment—which is somehow nonstandard. If I tell Mary something, I establish myself, then Mary-as-experiencer/joint-agent. But if I get cross with Mary, I think I (i) get cross, (ii) separately—*direct my behaviour towards Mary*.

Now consider another use of *with*:

(11.34) John squashed the bug with his thumb.

Here again, the PP is describing something happening peripherally to the main SM sequence. Again, it certainly happens that the thumb is established as the effector (I presume some kind of environment or place)—but we don’t have to report this operation. If we do, it comes out as a PP.

So—maybe—there are two kinds of ‘interrupting SM operation’, that occur in a different medium from the ‘central episode-encoding medium’ (whatever that is). Some are obligatorily conveyed, because they describe operations that set up the initial context for the episode-denoting medium, or because they occur part-way through operations in this medium. Sometimes these are conveyed explicitly as PPs (e.g. *Le chocolat plaît à Jean*), sometimes they are conveyed obliquely as DPs (*Chocolate disgusted John*), sometimes they are completely implicit (e.g. *John loves chocolate*). But the point at which they attach is somewhat variable. Other obligatorily conveyed interrupting SM operations that occur partway through operations in the episode-denoting medium are reported by PPs with a somewhat less variable position: for instance *John went to the door*; *John put the cup on the table*. The other type of interrupting SM operation is the kind that doesn’t have to be reported. These surface as PP adjuncts. For instance, *John squashed the bug with his thumb*.

11.5.2 The 'complete' model of SM routines

In the complete model, I have to envisage a number of SM media, each of which allow SM routines to be executed, stored in working memory, and replayed. These routines are partly autonomous, and partly connected together in various ways.

11.5.2.0.1 The different SM media So far I can think of three.

1. The episode-denoting medium. (For storing episodes that update the reference time.)

2. The individual-denoting medium. (For storing attentional routines establishing individuals and groups. And maybe also predications.)

3. Perhaps: an environment-update-denoting medium. (A medium representing updates to the current WM environment.)[31]

I'm not sure why environment updates would happen in their own medium. Sometimes they're legitimate.

11.6 Objects and environments: a recap from Part 2

A useful idea to recall from Part 2 is that there's a whole cognitive modality specialised for recognising environments: see Section **??**. This includes representations of environments as wholes (see in particular the parahippocampal place area and retrosplenial cortex) and representations of environments as maps of places (see in particular the hippocampus[32]). I'm trying to think of a cognitive medium distinct from the episode-encoding and individual-encoding media that implements actions of attention to environments—and an associated WM medium that *stores*, and is able to *replay*, such operations.

Obviously there's a link between environment-encoding media and object-encoding media: the maps of places within an environment-encoding medium are related to—if not identical to—*saliency maps*. Objects are found at places in environments.

Also recall from Part 2 that when you attend to an agent as an individual, you *automatically* activate a set of motor maps centred on this individual's various motor systems, that represent the agent's peripersonal space. I think this means that the current environment representation is updated. But this update happens within normal processing in the episode-perception modality: perhaps the new environment is represented as part of the 'new context' established after attending to an agent at AgrSP (i.e. in syntactic terms, the new motor environment is pointed to by [Comp,AgrS]). This is the medium in which objects in the agent's perispace compete for selection as targets of motor actions. So: maybe *this* shift-of-environments is naturally described within the episode-perception

[31] Maybe it could also be considered a *mode-change-denoting* medium? There's certainly a link between environment-resetting and mode-changing. For instance, if I establish action execution mode, this sets the environment to be my motor environment; if I establish external perception mode, this sets the environment to be the one which the attended object is in.

[32] But also motor maps, if the body is considered to be an environment.

medium (more like the motor action medium, maybe)—but the shift you need to do to *break out* of the motor environment and establish a new *allocentric* environment needs a nonstandard operation—or at least, an operation that's reported by a PP.

11.6.0.1 The environment in which internal episodes happen

In the light of the above discussion: I have to assume a special cognitive modality for representing agents' brains as environments, where different stimuli can arise, and cause different feelings. My suggestion is as follows.

1. The medium in which possible stimuli compete and are selected is the semantic medium representing all possible object tokens and types. We abstract completely away from spatial locations: assume that when we're 'contemplating' an object, the only relevant thing is the active semantic representation. This is an interesting saliency map, in the sense that objects are *already categorised* in this map! We are not attending to a map of uncategorised locations here: we are attending *directly* to *semantic objects*. There's a huge difference between the subject DPs in *John grabbed Mary* and *The chocolate delighted Mary*. In the former case, the observer attends to an external location, and activates this location; then, as a separate operation, the object classification system classifies the stimulus at this location and returns the individual 'Mary'.[33] In the latter case, the observer attends to a 'location' within the (very multidimensional) space of object categories: when the 'winning location' is found, it *just is* the semantic representation of the individual in the world referred to by 'the chocolate'.[34] Note: we still have to escape to the DP medium to *refer* to this individual in a sentence: but this is a *communicative* requirement, rather than a direct report of experience. I presume that WM individuals are indexed by identity as well as by location, so there's a way of adjoining a DP to the [Spec,AgrS] position in this case too.

Note: in the above account of saliency maps, we're trading on the fact that a saliency map is, at a technical level, just a neural region where alternative representations compete against one another and one is selected. I think this allows the space of semantic representations to be *construed* as an environment containing individuals, by some 'second-order representation' (to use Damasio's terminology).

2. The medium in which possible feelings compete and are selected is the anterior insula. Again, there's no indirection here: what are selected are not points in the world to direct a classifier to, but *semantic objects*. In this case, the medium representing these semantic objects is not a medium storing the output of a classifier directed at the external world, but a medium *classifying the internal state of the observer himself*. There's still

[33] Or the observer makes himself the agent. I think here again, classification is indirect: an allocentric location is again activated, and the observer objectively classifies the physical stimulus at this location, which happens to be his own body/self. The 'direct', or 'automatic', identification of self, is represented at the point in the routine associated with [Spec,VP], I think.

[34] If the sentence was *Chocolate delighted Mary*, we might be able to bypass the DP system altogether—or at least the referent-introducing part of DP (the SDP in Zamparelli's terms). Maybe this idea will be helpful in an account of the semantics of bare nouns / generics.

classification to be done: there are still clusters of pre-attentional feelings to be identified. But this classifier is more basic, in the sense that (a) the space of possible outputs is much smaller (there's a reasonably small number of possible feelings, because they relate to homeostasis, as Damasio says) and (b) they directly represent the state of the agent's *body*: they represent what's *in* the organism, rather than what's outside.

11.6.0.2 The operation denoted by *in John*

My suggestion (as already mentioned) is that the operation denoted by *in John* is the operation that establishes internal perception mode. This operation sets the 'current environment' to be John's mental processes. In implementation terms, as proposed above, the operation sets the active saliency map to be *John's semantic classification system.*[35]

Recall: my question was: what's the operation that establishes John 'as an environment'? I think the above discussion answers this question: John's mental processes are represented *as a saliency map.* But a special kind, in the sense that salient entities that compete, and can be selected, are semantic representations already.

11.6.0.3 Indexing to mental environments

To return to my discussion about PPs: I was suggesting that PPs always denote operations re-establishing the current environment. Clearly there are lots of different neural media that can function as environments. What do PPs all have in common?

One possibility is that it's something to do with LTM representations and indexing. I'll consider that possibility in this section.

I suggest that the establishment of a new environment is always represented in the same way *in LTM.* As discussed in Part 2 (Section **??**), my proposal is that each LTM individual is associated 1:1 with an LTM environment. Actually, if the LTM individual represents an agent, agents, there are probably several different environments, one for the agent's physical sub-parts (maybe including the motor system) and one for the agent's feelings.[36]

To show how this might work: if I activate an LTM individual representing John, then execute 'establish-as-feeling-agent', this activates an LTM environment linked to John,[37] which *modulates* associations between pre-attentional feelings and semantic stimuli, so they reflect associations identified *in John.* (Critically, the observer still uses *his own* brain to represent these. So he's establishing John through a process of *empathy* of some kind.) In the resulting state, semantic stimuli compete to be selected partly on the basis of their emotional associations. If that's the only source of competition, we pick the stimulus with the highest emotional valence of all (e.g. 'football'). But there may also be a focussing element of some kind: for instance, 'chocolate'—creating the implicit question 'As for the

[35] Note this doesn't just represent physical individuals: it can also represent situations, episodes, etc: for instance *Mary's dancing disgusted John.* (But they're all reported as DPs, mind you.)

[36] And probably others representing the agent's beliefs, desires, etc. But these aren't the current focus, because they're not expressed in PPs.

[37] And only to John? Probably not; most of the time our ascriptions of mental states are done *partly* on an individual basis, and *partly* by virtue of an individual's membership of one or more general classes.

stimulus ‘chocolate’, what is John’s feeling towards this stimulus?’ In either case, having selected a stimulus, we identify the most active feeling, and select that; that gives us the ‘love’ part. The important idea is that in either case, what the observer is doing is inspecting information stored in LTM about the feelings of other agents towards various stimuli, learned in the course of observing these agents. The sentence we might get is *John loves [*$_{TH}$*football]*, or *John [*$_{F}$*loves] [*$_{TH}$*football]*.

Another example might be: say I activate the LTM individual representing John, then execute ‘establish-as-physical-environment’. I think that operation is denoted by the word *have*, but a location within the new environment will also be established: and that’s denoted by a PP. So we might get, for instance, *John has a scar on his cheek*. The location is sometimes optional, as in this case. Again, the idea is that the observer *perceives* this fact by establishing John as a physical environment, and then records it in LTM by activating the LTM (physical) environment *associated* with the LTM individual ‘John’, and then evoking a set of locations within John, which are salient through having different contents than a prototypical person of John’s category. There’s no *in John* here, but there’s *on John’s cheek*, which is somewhat similar.

11.7 Back to *à Jean*

None of the above ideas about PPs relates directly to the question of what operation is signalled by *à Jean* in *Le chocolat plaît <u>à Jean</u>*. The last point that links to this is the idea that Jean is being established as an environment, represented as a set of places—and that these places are semantic representations directly in Jean’s head. Clearly, we are representing Jean *within the SM system that represents environments*. I need to think again about what this system is doing, in general, and how Jean-as-an-environment can be established.

As already discussed in Section 11.4.0.2, in order to establish Jean as an ‘internal environment’ (which is the same as ‘establishing feeling perception mode’), I first have to establish Jean as an individual: this goes both for the case where Jean is an external agent and for where I’m Jean myself. At the level of SM routines, this is the case. (You have to *notice an emotion in the selected agent* in order to go into feeling perception mode mode. You have to attend to yourself in order to allow your own emotional saliency map to compete to be selected. You have to attend to someone else in order to recognise that they’re having enough emotions to warrant going into feeling perception mode. Feelings are represented in a mirror area, so they have to be deictically referred to the currently-established agent.) It’s also the case in relation to a model of LTM:[38] in LTM we have to represent my feelings as mine, and John’s feelings as John’s, and this requires the LTM individual denoting me (or John) to be active to *modulate* the learned associations between semantic stimuli and feelings. So the operation I need to think about is the one that establishes attention on an individual agent, and then *re-establishes* this same agent *as a (mental) environment*.

[38] See previous subsubsection ‘indexing to mental environments’.

11.7.1 The spatial structure of a mental environment

If an agent is established as an environment, an interesting question is: what's the *spatial structure* of this environment? In a physical locomotion environment, the structure is given by a set of neurons in the hipppcampus; the topographical structure of the map they describe is determined by representations of the environment's boundaries. In a mental environment, the set of 'places' are in a different medium: for instance the medium that represents types of external stimuli, and the medium that represents types of feeling. I'm not sure this structure ever changes. (Although it's vaguely possible that semantic proximity relations between concepts are different for different agents, and that this can be represented using something analogous to boundaries.) For the moment, I'll just assume that

11.7.2 'Transitioning' into a mental environment

If I establish an object like a table as an environment *during an action*, I transition neatly into it, from some adjacent or local environment. I can only do actions involving environments I transition to locally, I think. But I can also establish an environment in much more discontinuous ways. In memory mode, I can be *reminded* of a very distant spatial environment. In feeling-perception mode, I can be drawn to contemplate a distant environment, because of its strong emotional connotations. My question for this section: when I establish John as a mental environment, what kind of shift is this? Must John be attended to as an object first? Must John be in my current environment, in the way he must be if I decide to put something on him?

Certainly he can be present before me. If I decide to attend to John, and then notice that he's having feelings, I can go into feeling-perception mode right then. But he doesn't have to be present before me. I can entertain the proposition 'John likes cheese' even if he's not here, because it's part of my LTM representation of this agent. Say I activate the LTM individual 'John', and subtract the representation of the prototypical object of his type (say the prototypical Kiwi man). It may be that there are enough feelings collectively activated that I enter feeling perception mode and activate the stimulus 'chocolate', then the feeling 'like'.

I think I need to attend to John as an object first even if he's not present. So: perhaps the minimal establishment of an agent as an environment involves establishing him/her as an object, then activating the 'associated' LTM environment.

11.8 Experiencing feelings in the WM episode system

[Not sure where this one fits in..]

I like the idea that the WM episode system encodes some sort of preset, fixed sequence of coordinate system transformations. Consider a transitive experiencer-stimulus sentence:

(11.35) I loved the chocolate.

In this case I attend to myself, then—in an operation *invisible* to the WM episode system—I establish feeling perception mode. At this point I have a *choice*. I can *retain* the operations stored so far in the WM episode buffer (i.e. reference to myself as agent), and continue adding operations executed in the new mode *into this same list*: in which case the buffer will end up holding operations executed in a *mixture of two modes*. Or I can *clear* the WM episode buffer and start again.[39] In the former case I get *I loved the chocolate*. In the latter case, I get *The chocolate delighted me*. The critical thing is a mechanism, *triggered by a mode-changing operation*—and apparently *optional*—that inhibits the contents of the currently active WM medium.

If I *don't* inhibit the representation of me 'as agent' in the WM episode system when I enter feeling perception mode, then the first operation occurring *in* this newly established mode (activation of the semantic stimulus 'the chocolate') is forced to be represented 'as the object of a forthcoming motor action', just like a cup would be. And the *next* operation occurring in feeling perception mode (activation of a feeling) is forced to be represented 'as a motor action' (on this 'object'). To me this says something useful about the *nature* of the WM episode system.

If I *do* inhibit the representation of me 'as agent' in the WM system when I enter feeling perception mode, then the first operation occurring in this mode is the first operation in the whole medium. Therefore it's forced to be represented 'as an agent'. The next operation is—I presume—activation of the feeling. And the final operation is something like 'cause-to-arise'.

Again, to me this says something very useful about the nature of the WM episode system. If we're using the WM episode system to report the latter process of identifying the cause of a mental event, it seems a very different kind of thing from a normal action. Maybe it even happens in its own WM medium? I've already discussed concepts of causality in Chapter 9—this is just a reference back to that.

[39] Clearing the WM episode buffer is something that's often associated with changing mode—for instance, it happens when you enter verbal mode. (Or even entering 'desire mode'. Consider *John$_i$ wanted [pro$_i$ to sing]*. Here we enter a new mode, *and also inhibit the material accumulated so far in the buffer*, so that the complement clause has no subject, and no subject agreement features.)

Part III

A model of utterances and sentence processing

Chapter 12

Representation of communicative actions, dialogue participants and dialogue states

12.1 Planning and representation of communicative actions

I previously had the idea that when I tell a person P something, I (i) make myself the agent; (ii) attend to P; (iii) activate the 'talk' action, which has the side-effect of (a) clearing the current WM episode, so that a new winner is chosen and (b) entering verbal mode and simulation mode. That's almost right—but it provides no indication of how I decide *what to say* to the person P.

A new idea is that there are two modalities in which I can attend to P. I can attend to P as a physical object, in which case I will represent her motor affordances for me. But I can also attend to P as a sentient agent, in a separate medium. In this medium, attending to an agent causes the candidate WM episodes buffer to represent something new: the things which might usefully be *communicated* to P. (Or more accurately, the things which *I might want* to communicate to P.) If one of these is sufficiently good, it will be selected, and I will execute a 'talk' action, to P. (The idea is that having attended to P as an interlocutor, we have already determined that the action, if any, will be 'talk'—or at any rate, some form of communication.[1] So we no longer need the WM episodes buffer to represent alternative actions of our own, and we can free it up to represent alternative possible propositions to be communicated.)

In this scheme, there's a new medium for representing agents: they're represented as modulations on the distributions in the candidate WM episodes buffer. In fact, agents are also represented as modulations on the distributions in the candidate WM episodes

[1] The exact type of communication—whispering, shouting, telling, showing, etc—might still have to be selected, in the same way as after we decide we are going to do a locomotion action, we still have to decide whether to walk, run, crawl, drive, etc.

buffer. But the two types of modulation are a little different. In the case of the agent, the modulations indicate good plans for the agent. In the case of the interlocutor, the modulations indicate the things that it would be good for the patient to know.[2]

Probably these two modulations have to apply in sequence. Because the interlocutor modulations represent the things it would be good *for the agent* if the interlocutor knew.

12.1.1 The interlocutor as observed agent?

I quite like the idea that the interlocutor is similar to an observed agent. The only difference is: when we establish the interlocutor, we have already established the speaker first. The speaker is an agent. Say the initial context represents the episodes that are possible, or likely, in the current situation. When I establish myself as the agent, I filter this distribution to identify the things that are possible and also *desired by me*. When I then attend to an interlocutor, it's a little like attending to an external agent. In fact what's happening is I'm creating a medium for representing *joint actions*: potential things that I want to do and that the interlocutor will go along with. But as the initial agent, I'm still in the driving seat I think.

12.1.2 An example

Take an example. The agent is a baby, and is in a situation where there's a cup that's too far from him to reach, and also another person (Mummy). The baby first makes himself the agent. This generates a particular pattern of activity in the candidate targets medium (say representing Mummy, because she's reachable, but not the cup, because it's out of reach). It also generates a pattern of activity in a separate medium, the candidate interlocutors medium, representing Mummy (but again, not the cup, because the cup's not sentient). The baby has to choose one of these media—he can't choose both. If he chooses the candidate targets medium, he will then choose Mummy in this medium, and will then select some physical action to perform on Mummy. If he chooses the candidate interlocutors medium, he will again choose Mummy. But this time the representation activated is of Mummy as a receiver of potential messages from the baby. There are several things the baby might communicate to Mummy; which of these are beneficial will have been learned through reinforcement. Some of these are things that Mummy can do (i.e. episodes in which Mummy is the agent). Some of these are propositions which if communicated to Mummy will get a rewarding reaction from her. Say the baby picks the episode in which Mummy gives him the cup. This episode is now the thing that's rehearsed in verbal mode.

I like the idea that the candidate WM episodes buffer is a deictic representation, whose content can change from moment to moment depending on the observer's current focus of attention.

[2] We might say that the original distribution of WM episodes represents the episodes that are *possible*, and the distribution after modulation by the agent representation represents the possible episodes that are also *desired* by the selected agent.

12.2 Representation of the interlocutor

The representation of the interlocutor focusses on things like: (i) what she *can* do (ii) what she *wants* to do (as it's easier to persuade people to do things they want to do); (iii) what she's interested in (as she's likely to give you a reward if you tell her something she's interested in); (iv) what she currently knows (as it's not a good idea to tell her something she already knows).

Perhaps we include representations of what she knows and we would *like* to know—which will surface as questions.

Technically, an interlocutor can perhaps be represented as a localist unit fully connected to the candidate WM episodes layer, so that each interlocutor can generate a full distribution of candidate episodes.[3]

12.3 Representations of content without interlocutors

I might want to give *someone* an instruction, nor caring who, or to ask *someone* a question, not caring who. In this case, the 'fixed' issue concerns the content, and the open issue concerns the interlocutor. In the above cases, the fixed issue concerns the interlocutor, and the open issue concerns the content. These are two extremes; I expect there can be in-between cases where interlocutor and content are selected in parallel.

Even if the content is fixed before the interlocutor, I assume the first operation must still be to establish an interlocutor—it'll just be a very underspecified one. This is necessary in order to change the interpretation of the candidate WM episodes buffer (so it represents the contents of possible utterances rather than possible episodes). In this case, there must be something in the candidate WM episodes buffer prior to establishing an interlocutor that strongly pushes for a communicative utterance to be made. What might this be? E.g. if my main goal is to get X and I can't find X, this would be reason to ask someone where X is. It may also be that there's an explicit rule: if you can't find X, ask someone where X is. I think this kind of rule can be implemented in the mechanism that updates the candidate WM episodes buffer (i.e. updates the current situation rep).

12.4 Representation of jointly entertained information

In a communicative action, the agent first establishes an interlocutor, (which means he gets her attention). At this point, any semantic content the agent evokes will also be evoked by

[3] This idea will also be helpful in modelling updates to the situation. We can imagine the candidate WM episodes medium as a recurrent SOM, that also gets input from a medium isomorphic to the candidate WM episodes buffer that represents the episode just experienced. This way we can learn a function from (i) a just-experienced episode and (ii) the situation in which it occurred to (iii) the next situation (i.e. an updated distribution of WM episodes).

the interlocutor (if things go well). This is because after establishing an interlocutor, the agent enters a special mode where the semantic content he activates has overt behavioural reflexes.

A simple example is *John showed Mary the flower*. John first makes himself the agent, and then establishes Mary as an interlocutor. He does this by catching establishing joint attention with her (perhaps this involves a term of address or a sound of some kind) and indicating that he has something to communicate (which is could be done by an eager expression, eyes wide, etc). Mary probably indicates by focussing her attention on John, so John has some indication he has succeeded in getting her attention. (These behaviours are quite subtle.) At this point, John attends to the thing he wants Mary to attend to (the flower). He does so in a sufficiently public way that Mary can tell what it is, and can do likewise. This may be just eye movements, but the simplest action is a pointing action. That could start off as a reach to an out-of-reach object. Later on, a word might be used, or might accompany the pointing action.

A more complex example is *John told Mary that the dog was big*. John establishes Mary as interlocutor in the same way, so she's guaranteed to activate whatever semantic representation he activates. (We know this because (a) she's attending to him, and (b) he's in a mode where the content he activates has behavioural reflexes. In this case, a talk action is done.) The key idea is: **anything that's selected after the interlocutor is established is guaranteed to be jointly entertained by speaker and interlocutor**.[4]

12.4.1 Dialogue referents

I've been thinking so far mainly of WM episodes. But another idea is that when an interlocutor is established, the agent also changes his representations of WM individuals, so that the only active ones are those that are (a) of relevance to his interaction with this interlocutor, and (b) *known by the interlocutor*. This would provide an account of the 'referents' in a traditional dialogue model (e.g. DRT): the things that the hearer knows about as well as the speaker.

On this way of thinking, the WM system (WM episodes and WM individuals) can serve as the neural medium representing the 'dialogue context', as envisioned within linguistic dialogue theories. Very roughly: the WM individuals system represents the referents, and the WM episodes system represents the 'conditions' (in DRT parlance).

That's a bit of an approximation. After every episode is presented to the hearer, the hearer replays it to episodic LTM, so the DRS is partly stored in episodic LTM. So the *selected* WM episode represents the new condition to be stored in the DRS. The candidate WM episodes buffer should be thought of as representing 'the current situation': i.e. a particular distribution of possible/desired episodes.

[4] Of course the interlocutor can indicate she failed to hear or understand, but we can think about that separately.

12.4.2 Introducing hearer-new referents

Say I want to tell the hearer that John chased a particular dog, that is known to me but not to the hearer. The utterance must introduce the dog as new to the hearer. But before the utterance is produced, I must have a way of planning it, and the dog must be present at this time.

One possiblity is that during the planning process (i.e. the process of selecting an episode to communicate) I work with my own set of WM individuals, but once the episode is selected, I filter out the referents unknown to the interlocutor

12.4.3 Updates

If it's true that within communicative actions, WM individuals and episodes hold representations of 'the current discourse context', then there should be a sensible notion of an *update* to a discourse context. E.g. if the utterance *John chased a dog* is made, the common ground should update to include a representation of the newly-introduced dog. Certainly, the distribution of candidate WM episodes should update in the right way, to represent the intangible knowledge-based updates to the discourse context that determine what's likely to happen next (and help resolve ambiguity in language). But the set of active WM individuals should also update in the right way.

12.5 Enduring representations of speaker and hearer

The participants in the episode described by the utterance don't have to include the speaker and hearer—they can be anyone. Nevertheless they *can* be the speaker or hearer. When they are, they are realised linguistically as 1st- or 2nd-person pronouns (e.g. *I*, *you*). So there must be representations of the speaker and hearer that endure right through the communicative action.

One possibility is that the plan-based representations of the agent and interlocutor endure in the right way, and serve as the denotations of 1st/2nd person pronouns. When the speaker establishes himself as agent, he activates his own set of desires, goals, plans; I think this representation has to endure right through the communicative utterance. When the speaker establishes the interlocutor, he activates a representation of the interlocutor, which must likewise endure through the communicative utterance. Maybe the representation of the agent is realised as *I* and the representation of the interlocutor is realised as *you*.

How would these representations of speaker and hearer be activated during verbal-mode rehearsal of the content of an utterance? I assume that the content is rehearsed in the normal way: agent and patient are indexes into the WM individuals system. Maybe speaker and hearer are a special kind of WM individual.

12.5.1 WM individuals associated with first- and second-person referents

I represent 'myself' in a variety of modalities. When I decide to act, I generate a representation of myself as a collection of plans and expectations in the candidate WM episodes buffer. (This is a representation of myself *as an agent.*) But I also represent myself objectively, as an object at a location in the world. I suggest that when I decide to act, there's circuitry that associates this operation with a representation of my own location (in the local environment), and also with a representation of myself as a complex of properties (including the winning property assumbly 'man').[5] This is partly hard-wired: I assume there are two functions for activating locations in the cognitive map: one that delivers my own location, and one that delivers an external location to which I'm attending (see Part 2, self- and other-location functions). But I may have to learn that attending-to-self should trigger the self-location function. All of this relates to infants' development of an objective conception of self, as evidenced in the mirror self-recognition task and similar tasks.

The above cross-modal associative mechanism provides a means whereby my representation of myself *as an agent* can index to a representation of myself as an object *in the WM individuals system.* Deciding to act basically accesses a WM individual, because it activates a location and a property complex. This is exactly what I propose happens in references to external objects. In that case I (a) decide to perceive rather than to act, and simultaneously (b) attend to a particular external location, categorising the entity at that location as an agent. The decision to perceive activates the external object location circuit, so that the location of the perceived external object is activated in the cognitive map. The location and type jointly index a WM individual representation. So both for a first-person referent and for an external referent, representations in the WM episode system index, or point to, WM individuals.

One question that arises concerns 2nd-person referents—to be concrete, references to the interlocutor. I presume this also accesses a WM individual—but there must be something special about it, distinguishing it from 3rd-person references to objects external to both speaker and hearer. I think to get at this distinction I need to step back a bit and think about the contexts in which the words *I* and *you* get used.

12.5.1.1 Speech situations

When I decide to act, and thus attend to myself as an agent, there will be no activation of the word *I*. Words are only activated when I'm in verbal mode, and I haven't decided to enter this mode yet. Likewise, when I decide to attend to Mary as an interlocutor, there will be no activation of the word *you*: the decision to enter verbal mode still hasn't been made. It's only after having attended to an interlocutor that I decide to enter verbal mode. At this point, as discussed above, I clear the active WM episode (somehow retaining plan-based representations of myself as agent and Mary as interlocutor) and activate a representation of the content I want to convey to Mary.

[5] And possibly also with the cardinality 'singular'. But I'm not sure how that happens.

Let's say what I want to convey is an action I recently performed, in which I grabbed a cup. (Don't worry for the moment why this episode is chosen!) The WM episode representing this action is selected, and rehearsed, in simulation, in verbal mode. The first operation is the decision to act. I'm simulating the operation, not doing it, so I don't *actually* decide to act. (I'm already acting, in fact.) The consequence of the operation is activation of two representations. The decision to act activates a representation of self that is *by definition the agent that is speaking.* The linguistic reflex of this representation (I assume) is the word *I*. The decision to act also activates a representation of the agent as an object. The linguistic reflex of this representation is (let's say) a proper name. So the speaker's linguistic system has a choice of two words: *I* or *Ali*. Early in development, it picks the proper name. Later on, perhaps when the agent's plan-based representation of himself becomes sufficiently different from his plan-based representations of other agents, he can choose the word *I*.

Now say the speaker wants to express an episode in which *the addressee* grabs a cup. The addressee is represented as an external object: the thing that was attended to in the sensorimotor routine. What the speaker simulates is (a) the decision to perceive; and (b) the representation of a particular external agent. (Since he's simulating, he doesn't *actually* decide to perceive; he continues to execute (a verbal action). Nonetheless, the activated representation can activate a WM individual, which allows expression e.g. of a name (*Mary*) or of a type (*lady*). But in addition, there are cross-modal associations between the WM individual *and the speaker's plan-based representation of the interlocutor*, so that this representation is re-activated. This allows production of the word *you*.

Notice that in the case of *I*, the representation of the speaker as agent precedes the representation of the speaker as external object—while in the case of *you*, the representation of the hearer as object precedes the representation of the hearer as interlocutor. But in either case, it's the same cross-modal associations between representations of the agent and interlocutor in the planning medium and in the object-representation medium that underlie representations of *I* and *you*.

12.5.1.2 The swappability of *I* and *you*

The above account of how *I* and *you* refer back to representations of the speaker and hearer enables a nice story about how the referents of these words depend on whether you're speaking or hearing. When you're speaking, the speaker is you, so *I* refers to yourself. When you're hearing, the speaker is someone else, so *I* refers to this other person. And vice versa for referencs to *you*.

12.5.1.3 3rd-person referents

When I attend to an external individual while simulating an episode in verbal mode, there's obviously a big difference between attending to the interlocutor and attending to something else. At bottom, they're both references to external locations. But if the external location happens to activate my plan-based representation of the interlocutor, this is a strongly preferred route for word production: if I activate the interlocutor, I should certainly say *you*, and I should only use 3rd person if I *don't* activate the interlocutor. I can't quite work out why that is. But I think it might turn on my reasons for picking this interlocutor in the first place. I chose to talk to this individual because of the things she can usefully be told; these are very often going to relate to her (especially if she is the agent of a requested action). The *selection* of the content to be communicated thus draws heavily on the plan-based representation that functions as the referent for *you*. I still haven't quite worked this out, though.

12.5.1.4 The pronominality of *I* and *you*

Another nice feature of the above account is that the pronouns *I* and *you* are always references back to recently established WM individuals, just like 3rd-person pronouns. Before you produce the words *I* or *you*, you must have (a) decided to act (and therefore established the referent for *I*) and (b) decided on an interlocutor (and therefore established the referent for you).

12.5.1.5 Plan-based representations in verbal mode

After I've selected myself as agent, and Mary as addressee, and decided to talk, I clear the current WM episode to make way for a new one. Let's say I want to communicate an episode that itself involved a speech event: John (a third party) told me [to sing]. This is an interesting case, because the representation of John as speaker can't overwrite my representation of myself as the actual speaker of the utterance: I in fact feature as the *addressee* of the nested speech event (in which role I must express myself with the pronoun *me*).

Note that when I report to Mary an event in which John says something to me, there are *two* clearings of the episodic buffer. The first happens after I establish myself as agent and Mary as addressee and decide to talk: the episodic buffer representing the physical talk action is cleared to make way for its content. But expressing the content also requires a sequence of two representations in the episodic buffer: first John's talk action, featuring me as addressee, and then the content of this action (the imperative *sing*). There are two questions about the nested talk action. Firstly, as already mentioned, how can we represent John as speaker and me as hearer of this nested utterance, without overwriting the representation of me as speaker of the *actual* utterance? Secondly, when I clear the episodic buffer for the second time, where does the content of the utterance come from? In Section **??** I proposed that the content of a communicative action is the second-most active item in the candidate episodes buffer, that becomes most active when the talk action itself is inhibited. That seems okay for top-level communicative actions dealing with the current situation (e.g. if I tell Mary to give me a cup, or point out that there's a person at the door) but not if I'm telling her about something unrelated to the current situation. If I tell Mary that John told me to sing, I'm telling her a *story*: the mechanism for evoking the content of the story must be different, I think.

I'm going to assume that in my utterance I'm describing a situation recalled from *episodic memory*, i.e. from long-term memory. First, a recap about how this works. Episodic memory stores sequences of context representations, linked by hippocampal representations of episodes that move from one situation to the next. (I'll use 'hippocampal' in lieu of 'cortical', though these things are eventually consolidated in cortex.) Each context representation points to a particular distribution of activity in the candidate WM episodes and individuals buffers; these can be used to help 'reconstruct' what happened next—though I also store what *actually* happened next. So what I do after clearing the current WM episode representing myself as speaker and Mary as interlocutor is to activate a LTM context, creating an active WM situation, and then read out the episode that actually happened next, in which John told me to sing. Crucially, the representations of 'John' and 'me' [Ali] are quite independent of my plan-based representations of myself as speaker and Mary as hearer. They don't rub out these plan-based representations—so if the content of my utterance happens to refer to the individuals 'Ali' and 'Mary', they could be expressed with 1st and 2nd-person pronouns. There still *is* an active WM situation, but it's established from representations of situations in long-term episodic memory. This addresses the first question.

As to the second question: I suggest that after the 'second' clearing of the episodic buffer, the content of John's utterance comes from episodic LTM. Recall that episodes are recalled from LTM as sequences, that re-create WM episodes; I suggest that the sequence is just a bit longer in the case of a reported speech action, and that the sequence replayed from LTM includes the operation of self-inhibiting, followed by another sequence conveying the desired 'sing' episode.

Finally, it's interesting to ask how I get started on my story—i.e. how I decide to activate the story-starting LTM context when I inhibit the episode in which I talk to Mary. I guess I know she'll like to hear this story. It's a story about John: perhaps Mary is interested in John (and I like to tell her about topics she's interested in). Perhaps it's also a funny story, and Mary likes funny stories. In general, we are reminded about things in memory by things in our current situation. This could be direct: for instance, when I see a dog, I might recall stories about the dog. But it could also be indirect: if I'm establishing Mary *as an addressee*, I might recall stories about things that Mary is interested in. Perhaps I consider several things to say, one by one, and in each case, consider (a) whether she already knows it [which would rule it out]; (b) imagine how she would react.

How would the episode be identified as a story *about John*? One possibility is that John's participation in the talking event could be recorded by itself as an episode-level property: in this case it could occur to me as an interesting fact about John. But if it's a *story*, then it involves a whole sequence of events. It would have to be stored as a larger sequential unit in episodic memory (a little bit like a plan, I guess). In this case, there would have to be a way of indexing stories to the individuals, places, activities etc that feature in them.

12.6 Representation of joint actions?

Since the agent and interlocutor are both represented within the domain of desired episodes, we can imagine a modulation that represents the episodes that are desired by both parties. This allows us to start thinking about how the concept 'we' is represented.

I envisage an action that establishes the agent and the interlocutor as *joint agents*. This may involve (a) someone starting to help someone else (and their offer being accepted); (b) someone starting a joint action on someone else, as an invitation to a joint action (and the other person taking up the invitation)—e.g. one person starting a fight and the other one fighting back; (c) someone verbally proposing a joint action and getting a response.

None of these things have the flavour of a single attentional operation. Both participants are involved. There's an initiation and an acceptance. Maybe at the end of this process a joint planning medium is established. In each case, there's a shared goal (e.g. to move a piano, to fight). This goal has to be proposed by one participant and accepted by the other. So in a sense, the goal is active before the

Chapter 13

Sentence generation?

13.0.1 A general architecture for sentence generation

The input to the process is a *set* of stored, replayable SM routines, with various properties.

Firstly, there is some *redundancy* within the collection of routines. Not every element of every routine needs to be expressed in surface language—only enough to allow a hearer to reliably recover the set of routines. (But I think the hearer does reproduce the whole set as a result of the parsing process.)

Secondly, the routines are indexed to one another in various ways. So if you start with one routine,

My idea is that an episode representation in WM consists in a whole set of redundant, inter-indexed stored SM routines—and that this whole set forms the input to the neural network sentence generator. This means that learning a language involves not just learning how to read out one *particular* type of SM sequence (e.g. one encoding a transitive action, like we have at present), but also learning the following:

1. For a given collection of SM routines, *which routine the rehearsal process should <u>start with</u>*.

2. For any given SM routine being rehearsed, *which opportunities to rehearse <u>indexed SM routines</u> (as* **embedded routines***) should be taken.*

13.0.2 Seeing

Consider *John saw the cup* (from John's perspective). Before I introduce a proposal about the SM routine that's being described here, I want to introduce Damasio's (1999) account of core consciousness in the scenario where an agent attends to a cup, because it provides a very nice framework for the SM routine that I propose.

13.0.2.1 Damasio's account of core consciousness of the cup

Damasio is interested in what happens when we consciously observe a cup, and thereby become conscious of ourselves, the cup, and some relationship between these two entities.

He also emphasises that this process involves a temporally extended routine: what he calls 'a narrative without words', that 'unfolds in time', and has 'a beginning, a middle, and an end'. Here's what Damasio says:

> We become conscious (. . .) when our organisms internally construct and internally exhibit a specific kind of wordless knowledge—that our organism has been changed by an object—and when such knowledge occurs along with the salient internal exhibit of an object. The simplest form in which this knowledge emerges is the feeling of knowing, and the enigma before us is summed up in the following question: By what sleight of hand is such knowledge gathered, and why does the knowledge first emerge in the form of a feeling?
>
> The specific answer I deduced is presented in the following hypothesis: *core consciousness emerges when the brain's representation devices generate an imaged, nonverbal account of how the organism's own state is affected by the organism's processing of an object, and when this process enhances the image of the causative object, thus placing it saliently in a spatial and temporal context.* The hypothesis outlines two component mechanisms: the generation of the imaged, nonverbal account of the object-organism relationship—which is the source of the sense of self in the act of knowing—and the enhancement of the images of an object.
>
> (Damasio, 1999:168–169)

Damasio proposes that the brain's most elementary pre-conscious representation of 'self' derives from its representation of the body as an *organism*, in the neural areas concerned with bodily homeostasis. These areas issue commands to the body (including the brain) that regulate its state and ensure its survival. The brain's most elementary pre-conscious representation of an external object like a cup is a 'first order neural pattern' in sensory and motor cortices. The wordless narrative outlined above describes a process involving the pre-conscious representations of the organism and external object, which creates what Damasio calls **core consciousness**, of both the self and the external object.[1]

13.0.2.2 A proposed SM routine associated with seeing the cup

John first attends to himself as agent. Then he attends to the cup. Note: in some sense he perceieves it—he could act on it—but in my model he's not done yet, because he hasn't represented the seeing action, or its effects. My idea is that in the scenario we're describing, the cup is perceptually salient enough—in fact, *visually* salient enough—to cause John to

[1] Damasio doesn't just assume that the brain creates 'wordless narratives' of its own sequential processing to represent the agent's simple actions of seeing external objects: he assumes this narrative-construction process is quite ubiquitous, and also operates on motor actions, and (I believe) actions of memory and planning too. In this sense, his model of second-order narrative construction is a lot like my model of WM-episode construction. Also in the intermediary role it is assumed to have been SM experience and language.

pre-emptively enter external perception mode. (Specifically, a version of this mode that emphasises visual perception.) I suggest this operation is the denotation of the main verb *see*. But now there are two other things that need to happen: these are the things that I think correlate well with Damasio's account as summarised above.

13.0.2.2.1 1. Re-attention to self I suggest that while John is executing the 'enter (visual) external perception mode' operation, he re-attends to himself 'as an animate agent', in the same way that he does when he's performing a normal motor action. What does it mean to attend to yourself 'as an animate seeing agent'? I suggest this is exactly the operation Damasio describes as John 'constructing and exhibiting the knowledge that his organism has been changed by an object', through which he obtains core consciousness of his self. The whole idea of the re-attention to self operation that's executed as a reafferent consequence of executing an action—any action—is that the self is represented as an animate agent, something that undergoes characteristic changes. In the case of entering external perception mode, these changes are changes in the organism's *mode* rather than in physical disposition of the body. (So I *think* what I'm saying is that 'cognitive mode' is a fundamental part of the organism's *state*, just like physical body disposition.) Damasio says that the brain generates a temporally extended, nonverbal *account* of the process whereby the organism's state is changed by the external object. I suggest that the framework for this account is a deictic, SM routine. It's nonverbal—but as Damasio also allows, it can be verbalised: there are words that denote the representations and processes that feature in the account. The word for the process by which the organism's state is changed by an external object is the first-person pronoun *I*.

In fact it's oversimplifying things to suggest that *I* denotes this one process. The word *I* (for a mature speaker, at least) denotes the speaker's *multimodal* conception of self, that combines self as experiencer, self as agent, self as rememberer, and so on. In fact my SM routine helps explain how this multimodal conception is formed. Recall that the representation of self 'as enterer of perception mode' is axiomatically constrained to be associated with the representation of self *as agent*, that's denoted by the initial action of entering action execution mode. So my account of SM routines, with its emphasis on reattention and the creation of multimodal representations, addresses an important question about Damasio's model.

13.0.2.2.2 2a. Re-attention to the cup When I grab a cup, the second representation of the target object 'cup' is generated in the modality of touch: it only arises when the grab action is *completed*. Is there a special new representation of the cup when I have finished 'executing' the 'action' of establishing (visual) perception mode 'on it'? Damasio certainly assumes so: he says the process by which the organism's state is changed (that provides core consciousness of the self) also 'enhances the image of the causative object'. What might correspond to this process in my model?

In my account of feelings, the second occurrence of the object is associated with the moment that a multimodal representation of the object is created, by associating the feeling it elicits in the agent with the perceptual representation of the object. I'm pretty sure that something like this happens in the case of seeing too. Here I suggest that *both* representations of the target object are visual, but come from different visual pathways. The first representation of the target object is generated in the dorsal (parietal/premotor) pathway: this computes the motor affordances of the object. When John enters external perception mode, he enables the *ventral* (inferotemporal) pathway, that computes 'semantic' representations of objects. When this pathway is enabled, and a stable representation is evoked in it, there's another, very interesting opportunity for learning a multimodal object representation. In the case of grabbing, the parietal/premotor perceptual representation of the target object in the motor system is axiomatically associated with the physical motor state generated as a consequent state of the grabbing action. In the case of *seeing*, I suggest the parietal/premotor perceptual representation of the target object in the motor system is axiomatically associated with the *inferotemporal* representation generated as a consequent state of the *seeing* action. The associative learning process is actually exactly the same—as it should be, in my model, given that *see* is a transitive verb.

13.0.2.2.3 2b. Reattention to the cup: creation of object representations in PFC Don't forget: the WM episodes system is in prefrontal cortex (PFC). The planned action of attention to a cup is a search goal in PFC: 'go and find an object of this type'.[2] Maybe the crossmodal association learned at the consequent state of a 'see' action doesn't link inferotemporal object representations with object representations in *parietal/premotor* cortex, but with object representations in *prefrontal* cortex. Maybe the important thing about entering external perception mode (and thus enabling the IT pathway) is that it allows the prefrontal representation of the target object to encode semantic information about the object. Otherwise, it could only encode information about physical motor affordances activated in the parietal/premotor pathway.

Consider the model of simple transitive WM episodes for a moment. The agent attends to a target object (and gets a representation of this object), then does an action on this object, and in some cases receives a reward. If he gets a reward, he selects a unit in the candidate WM episodes buffer to represent the whole action, and links it to the *PFC* representation of the target object and the *PFC* representation of the action. Then in some future circumstance, if this unit is selected as the dominant plan, it will activate the PFC representation of the target object *as a search goal* and the PFC representation of the action category: if the search goal is achieved, the action will then be executed.[3] There are two points here. Firstly, when a SM operation is executed, it has to be represented *tonically*, in *PFC*, as well as in transient SM areas: otherwise there will be no PFC sequence to associate with a PFC unit. Secondly, the likelihood of getting a reward for a particular action done on a particular target object depends on the *semantic* category of the object, not just

[2] This goal is communicated to the WM objects system, which it must be remembered is *also* in PFC. All of the semantic WM systems are in PFC.

[3] During replay of an episode, e.g. for sentence generation, activating the PFC representation 'as a search goal' has *another* role, in *indexing an active WM object representation.* This is something to think about elsewhere.

on its physical motor affordances. So PFC representations of objects as they participate in motor plans need to make reference to semantic information: for instance, they must distinguish between apples and tennis balls. (These have the same grasp affordances, but different reward patterns: eating the *apple* is good, but eating the *ball* is bad.) We know that object representations in PFC can make arbitrarily fine-grained semantic distinctions, but place objects into categories that are relevant to the task at hand rather than reflecting arbitrary semantic properties (see e.g. Rainer *et al.*, 1998; Li *et al.*, 2007). So if there's no difference in task-related terms between a blue tennis ball and a red one (i.e. if there's no difference in their *reward* value), then PFC won't represent the difference at all. If after learning about the reward associated with doing the action 'eat' on an apple, the agent encounters an apple, the WM episode that becomes most active should be one that activates the planned motor action 'eat' rather than some action that brings reward when done on a tennis ball. If in a given situation, the agent activates the WM episode 'apple, eat', then the first item in this plan ('apple') should be activated as a search goal, in a way that connects with a lower-level search goal in IT, resulting in a search for an actual apple.

In summary, it's important that semantic information (originating from IT) gets into representations of goal target objects in the PFC-based WM episode system. Representations of target objects in this system have to be learned through some mechanism: I don't think I've yet considered what this mechanism is. I think it's distinct from the associative mechanism that learns affordance-based object representations in parietal/premotor cortex. That certainly has to do with reattention to the object in a reward context, but it relates specifically to parietal/premotor cortex. The learning that happens in the WM episode system is a bit different, I think. When you grasp a cup, you learn about the cup as a goal motor state, based on an internally generated reward of some kind ('the joy of touch' in the simplest case). But you *also* learn something relating to the goodness of doing *the whole action* in the current context. For instance, if in context C you bite an apple, you get a reward, but if you bite a tennis ball you get a punishment: there has to be a mechanism that positively associates the WM episode 'apple, bite' with context C, and negatively associates the WM episode 'tennis-ball, bite'. So that next time you're in C, the former plan dominates the latter in the candidate WM episodes buffer, and you're more likely to look for an apple.

At the end of the sentence we stop iterating through micro-contexts and update the macro-context. So the end of the sentence is when we do the piece of learning just described. This learning certainly involves the stored, replayable representation of the SM sequence that's just been executed. But note: the learning features the operations in the sequence *in the planning system* rather than as they transiently occurred: the operations that actually occurred, and their associated transient representations, are long gone.

I also need to find a syntactic analogue of the WM-episode-based learning of object representations. One possibility is that this relates to the fact that object DPs appear as the complement of V: this is the only place where they're not in a specifier position. So the whole DP is also a *context* in some sense: the consequent state of an action. Every other DP position is the reafferent representation generated by an operation, but the V complement is the consequent state brought about by this operation. My way of thinking of this fact before was that grabbing a cup is both a substantive action (that changes the world) *and* an epistemic action that's analogous to a saccade: the kind of thing that's signalled by an Agr head. What about in *John saw the cup*? I don't see how it can work in this case.

13.0.2.2.4 2c. Reattention to the cup: PFC object representations and WM situations The context is the second action of attention to the cup in *John saw the cup.* (Or *John tasted the chocolate*, for that matter.) I'm assuming the key thing is that this is signalled by the rightmost XP position in a transitive clause: the complement of V. The same position associated with tactile reward in *John grabbed the cup.*

Firstly, note that the lower object DP (at the V complement) raises to [Spec,AgrOP] in *John saw the cup* just as it does in *John grabbed the cup.* Assume that this raising links the IT-based representation of the cup (signalled by the V complement) with the affordance-based representation of the cup (signalled by [Spec,AgrOP]). That linking is somewhat *redundant*, because when the cup was first attended to within the episode-monitoring system as an affordance, it was also attended to within the individual-monitoring system as an IT-based representation: I presume the simultaneous activity of the IT-based and affordance-based representations at that early point permitted associations to be learned between the two. In this sense, the raising of the object in a sentence with the main verb 'see' is redundant—the learning that's being done here has already been done—but it still occurs. My suggestion is that the main learning associated with

The SM context described by the last XP in a clause describes the **consequent state** brought about as a result of the SM routine. This is a much larger thing than just the representation of an object. In this state, the world is different than before. We expect there to be a reward signal associated with this state that is distinguishable from the simple tactile signal that axiomatically teaches a child how to grasp a cup. I suggest that the reward signal is distinct from the mechanism that teaches cross-modal object representations (that's conveyed linguistically by DP-raising). When you grab a cup in a given context C, you don't just learn a cross-modal (visual/motor) representation of the cup: more centrally, you learn about the *value* of grabbing a cup in the context C, based on the goodness of the newly established context C_{new}. In network terms, you learn an association between the PFC-based representation of the cup-grabbing action as a whole (a unit in the candidate episodes buffer) and a representation of context C, whose strength is *proportional* to this value: negative if it's a bad outcome, positive if it's good. Value just has one dimension: nothing clever about feelings. The result of this kind of learning is a function that takes a current context C_0 and delivers a distribution of possible episodes in the candidate episodes buffer.[4]

Some contexts are intrinsically rewarding: for instance the context achieved as a consequent state of putting a biscuit in your mouth. Others are less so: for instance, the context achieved as a consequent state of picking up a biscuit. The value of this context is derived

[4] Martin has very nearly implemented this function in his WM model: it would be a quite trivial extension of what he already has.

from the intrinsically rewarding actions it *enables*: for instance the action of putting the held biscuit in your mouth. For concreteness, I'll assume that values are assigned to contexts by straightforward temporal difference learning (Sutton, 1988). Now every context has a value. So after every action episode A completed[5] by the agent, there will be a value. I'll call this value the **situation value**. (And contexts can also be called **situations**.) We might even say that the value is the sum of the activity of all the units in the candidate episodes buffer, because each unit is activated (or inhibited) in proportion to the value of the situation that will result if its associated action episode is executed. Or more precisely: we need to *add* this summed value (representing the value of *future* situations) to the value of any intrinsic reward arising right now—which will be experienced as a *feeling*, gradable on the pleasure-pain continuum. Anyway: the situation value of C_{new} may be more or less than the value we expected to obtain by doing the action episode A in C_0: the difference is used to strengthen or weaken the association between C_0 and A. That's a sketch of the temporal-difference learning process that happens at C_{new}.

Now: is there any way the syntactic structure of a sentence can be seen as referring to this process? If there is, by my SM interpretation of LF, it has to relate to the lowest XP in the spine of a clause. (In our examples, the complement of V.) This denotes the context where the WM situation update takes place, and we simultaneously have C_0, A and C_{new} in our sights. What is happening here? It's a DP position. But it doesn't have to be—for instance, it's not in an intransitive sentence.

Maybe the key thing is that the sentence *finishes when it does*. A sentence describes a SM routine, but there are natural stopping points in this routine, where learning happens. These stopping points coincide with *rewards*, from what I can make out. Why? Why do we get a reward when we finish grabbing a cup? Because it generates a tactile sensation? (Maybe that was the original reason?) Why is this a moment to update to a new WM situation? I suggest that there are actions that lead to actual rewards—maybe just rewards associated with tactile reattention, maybe substantive rewards like good feelings—and these bootstrap the system that identifies episode boundaries.[6]

I think it's a fair empirical generalisation to say that sentence boundaries are associated with reward states. Think about it: *John loved the chocolate*, *John put the chocolate in his mouth*, *The chocolate pleased John*, *The chocolate was delicious*. In *John grabbed the cup*, the axiomatic reward arrives in the position associated with the lowest XP in the clause. There's no reward in attending to onesself, or to attending to the cup as a target. There *is* a reward in *seeing* the cup. But in my analysis above, that reward arrives when the cup is *reattended to*.

[5] I like the idea that *completing* an episode has to do with re-attending to the object, or the agent. For instance, if you try to grab a cup but don't generate the expected haptic representation of a cup, you didn't complete the grab action. Or if you walk towards a target location but don't re-establish this location as your actual location, then you are still just walk*ing* towards the door (in the progressive aspect).

[6] Another possible mechanism for event segmentation has to do with prediction: this is an idea that has been pursued by Zacks and colleagues (see e.g. Zacks *et al.*, 2011). Knott (2012) suggests that inferring intentions in other agents probably has to work through a mechanism that makes successful predictions intrinsically rewarding, this mechanism may also end up relating to reward.

Now, consider what has to happen at the moment of reward. In the WM episode system there is a sequence of SM operations: each was activated when it was executed, and has remained active from then until the moment of reward. Now a unit in the candidate WM episodes buffer is chosen to represent the episode as a whole, based on simultaneous activity of all these PFC-based SM operations.

What do these PFC operations look like? The planned motor action is perhaps just a pointer to the premotor representation of the action. The target object is

13.0.2.2.5 Aside: Seeing and tasting The account I've given above of *John saw the cup* is the same as the one I gave in Section 11.3 of *John tasted the chocolate*. The only difference is the perceptual modality. But this makes me think that the re-attention to the chocolate in the taste modality isn't specifically about adding *affective* properties to a representation of the chocolate: that's something better associated with sentences containing words like *loved* or *pleased* or *delicious* (e.g. *John loved the chocolate*). Rather, tasting an object involves making cross-modal associations between the motor representation of the chocolate (whatever that is) and a gustatory representation. This is distinct from the affective representation: I can categorise a food by its taste, whether I like the taste, or dislike it, or am neutral towards it.

13.0.2.2.6 Aside: Cross-modal representations within the episode and object systems Prior to having the above ideas, my assumption was that the WM episodes system couldn't access rich 'semantic' object categories, but was constrained to work with purely affordance-based motor representations. All of the rich semantic object representations were assumed to be computed within the *WM objects* system, whose operation is described by DPs. Obviously there are cross-modal associations between the affordance-based object representations of the WM episode system and the semantic object representations of the WM object system—but I had thought the only role these had was to index WM object representations to representations in the episode system. (Or in syntactic terms, to allow adjunction of DPs to positions in the structure of a clause.) This created a problem, because the WM episodes system is where whole episodes are selected. I was comfortable with the idea that episodes are selected based mainly on chacterisations relating to object types rather than tokens (e.g that we plan to grab a knife, and it doesn't matter exactly *which* knife) but it seemed overly constraining to assume that all action planning is done purely based on physical goal motor states. For instance, an apple and a tennis ball have very similar grasp affordances, but beyond grasping, the things you want to *do* with these two types of object are very different.

Now a solution presents itself: by actively perceiving (i.e. *seeing*) objects in action execution mode, an agent can learn cross-modal associations between physical affordance-based object representations and semantic object representations *purely within the WM episodes system*. (I think that these semantic object representations are in *prefrontal* cortex. I'm going to assume these are mainly generated from inputs in IT, rather than in parietal/premotor cortex, that mainly does physical motor affordances.) Having learned such associations, when an agent attends to an apple, he activates much more than a grasp affordance in parietal/premotor cortex: via the type representation in IT, he also activates a semantic representation of the apple in the WM episodes system in prefrontal cortex. If the motor action he ends up doing on the apple is rewarding (e.g. if he eats it), then the PFC-based WM episodes system will learn an episode representation *featuring the PFC-based semantic object representation*.

In the model I'm proposing, the prefrontal cortex is trained to

At the same time, there's still a difference between the semantic object representation computed within the WM episodes system and that computed within the WM objects system. The former is still not much more than a bare type: it's associated with number and gender, but in a 'flat' way: there's no representation of the sequence of attentional operations that deliver this information, no possibility of relative clauses. I think the type information is also somewhat coarser.

13.0.2.2.7 Summary In this section I have argued that the two components of the process posited by Damasio as the origin of our core consciousness of both ourselves and of external objects are precisely identified in my SM model of the action of 'seeing', construed as a special kind of motor action, whose consequent state has no external physical manifestation, but takes place entirely within the agent's mind. In fact, the SM routines framework that I have proposed helps shed light on Damasio's model. Firstly, it subsumes the action of seeing within the general class of transitive motor actions. The cross-modal learning of object representations that happens in seeing is organised in exactly the same way as the cross-modal learning that happens in grabbing: it's just that the affordance-based object representation is linked to a different modality (that of semantic inferotemporal representations, rather than motor states). And the notion of the agent 'as a dynamic entity' is also identical to the notion featuring in an account of physical motor actions. Secondly, there's a link to the syntax of natural language. For Damasio, the process of coming to experience the perception of an external object is a *nonverbal* one—but as he acknowledges, language does have words that refer to components of the process: in particular the word *I*. My account is in agreement with this idea, but goes further: I argue that the process Damasio refers to is described in detail *by the syntactic structure of sentences reporting direct experience of objects*: sentences like *John saw the cup*. Methodologically, this is hugely useful: it means that the syntactic analysis of language can provide an empirical window onto the process that Damasio describes. It needn't just be a matter of introspection, as it largely is for Damasio.[7] Finally, and related to this last point: Damasio's account of the self extends far beyond the 'core self' of which we become conscious during deliberate acts of perception. There is also the **acting self** and the **autobiographical self**. My SM model also provides a general framework within which these representations of self can be connected to the core 'feeling' self that's the focus of the above account. To begin with,

[7] Actually Damasio does hint at a very close relationship between the nonlinguistic 'narratives' the brain creates for itself and natural language. But he doesn't discuss this connection at the level of syntactic analysis.

my SM model of perceptual actions *directly* links the self as agent of motor actions to the 'feeling' self. Recall that in my model of transitive actions, there are two cross-modal object representations learned: one relates to the target object; the other relates to the agent. We have already discussed the one relating to the target object. The agent of a perceptual action is represented first as an initiator of intentional actions: the operation that generates this representation is the operation establishing action execution mode, and the reafferent representation it generates is of the self-as-agent.[8] But the agent of a perceptual action is *also* later represented as the agent whose internal state is changed by this action. My general SM model proposes that experiencing an action provides the agent with an opportunity to learn associations between these two representations of himself. So my account of a perceptual action directly links the self-as-agent to the self-as-experiencer. But in addition, because the self-as-agent can also execute *other* mode-setting operations, *additional* cross-modal representations can be indirectly learned. I will discuss these in more detail in Section 13.0.3.

13.0.3 An account of the Damasio's extended conception of self

.

We already know that when the observer enters action observation mode, he is attending to himself 'as an agent'. I've just suggested above that when the observer enters internal experience mode, he's attending to himself as an experiencer—as a being with feelings. A very interesting idea: maybe *every* mode the observer can enter provides another way in which he attends to himself?

If each mode is associated with a representation of self in a different modality, then cross-modal representations of self aren't just associated with reafferent SM stimuli (e.g. the self as dynamic agent in action execution mode) but with *whole SM modes.* This is a very powerful idea, I think.

I'll flesh out this idea by talking about action execution mode, external perception mode, and memory mode.

13.0.3.0.8 The self as agent There's something it feels like to decide to act.

13.0.3.0.9 The self as perceiver There's something it feels like to decide to perceive.

13.0.3.0.10 The autobiographical self In particular: maybe when the observer enters episodic memory mode, he is *also* in some sense attending to himself? As a being with a history, and an extended existence in time.

[8] I presume there is something it feels like to decide to act, and that the core self also experiences this, somehow. Talk about this later!

13.0.4 A feeling of rapture

In my account of *a line of soldiers*, *of* signals the establishment of a different classification scale, without a change to the attended location. What's happening in *a feeling of rapture*? This is interesting because in Section 11.3 I proposed that John's decision to attend to his feelings is distinct from his identification of a *particular* feeling (the dominant feeling associated with a particular stimulus). *A feeling of rapture* seems to describe a process whereby the feeling is first categorised simply as a feeling, and then, after attention is deployed in some new way, as rapture. The fact that the word *of* can be used here is support for my analysis, I think.

13.0.5 Evaluative adjectives

Evaluative adjectives like *beautiful* have special syntactic properties; for instance they're often claimed to be structurally 'high' in nominal phrases (e.g. we say *a beautiful red balloon* rather than *a red beautiful balloon*, or *un beau grand ballon* rather than *un grand beau ballon*.) In the above model, whenever you attend to an object, you may activate a feeling: if these activated feelings are expressed within the DP system, I suggest they come out as evaluative adjectives. This is another idea relevant to Marantz' proposal that words are initially category-neutral roots: the idea is that the nominal system has a way of relaying emotional associations using altogether separate words from those used in the clausal system.

If emotional adjectives are read from feelings associated with perceived individuals, we have to give an explanation as to why they're so structurally high. My suggestion is that they're high because evaluation is entirely separate from classification. Objects can elicit emotions before they are classified, so evaluative adjectives don't have to wait until classification. On the other hand, size adjectives arrive simultaneously with initial classification (establishing size is a side-effect of classification), and colour adjectives come strictly afterwards, because they involve property-level IOR.

Exactly how the hierarchical structure of a DP encodes the order of SM operations when they happen in separate channels still isn't clear. But ultimately the point is to create a representation of an individual in long-term memory. We certainly want to represent the emotional associations of the individual, if they are prominent.

13.0.6 Pylkkänen's argument-selection parameter

Pylkkänen also proposes another parameter of variation between languages, which relates to the constituent that a causative head can take as its argument. She identifies three possibilities.

13.0.6.0.11 Causatives taking roots as arguments In one pattern, a causative takes a **root** as an argument: an open-class semantic item that has not yet been identified as a noun or a verb. If a causative head combines with a root in the syntax, there is an associated semantic process of composition creating 'cause-to-X', but still only one verbal category, namely 'cause': this is used to explain why in English manner adverbs can only modify the causative action, not the caused episode (c.f. Fodor, 1970: *Bill awoke John grumpily* implies the Bill's action was done grumpily, and not that John's waking was done grumpily).

13.0.6.0.12 Causatives taking vPs as arguments In another pattern, the causative head takes a vP as an argument. (Finnish is like this, which explains why you can get unaccusative causatives, with a causative above vP but no external argument.) In languages like this, an adverb like *grumpily* in a causative construction modifies the caused action rather than the causing action.[9] Bemba is an example of such a language. Examples from Givon (1976), quoted in Pylkkänen (2002):

(13.1)
Naa-butwiish-ya Mwape ulubilo.
I.PAST-run-CAUSE Mwape fast
(i) 'I made Mwape *run quickly*'
(ii) *'I *quickly made* Mwape run'

However, Pylkkänen argues that not all adverbs modify vPs. **Agentive** adverbs like *quickly* and *grumpily* indicate the manner in which an action is performed, but other adverbs more plausibly indicate the manner in which the agent participates in the event: these are called **non-agentive** adverbs. Examples are *willingly* and *unwillingly.* In causative constructions in Bemba, these adverbs modify the causing action:

(13.2)
Naa-butwiish-ya mumuana ukriitemenwa.
I.PAST-run-CAUSE the boy willingly
(i) *'I made the boy *run willingly*'
(ii) 'I *willingly made* the boy run'

The structure here is supposed to be:

(13.3) [$_{voiceP}$ I [$_{vCauseP}$ made [$_{vP}$ run] willingly]]

This contrast between the scoping options for agentive and non-agentive adverbs in causative sentences provides good evidence that the external argument is attached in a different way (Pylkkänen, 2002). To confirm that Bemba causatives take vP as their argument rather than a root, Pylkkänen notes that verbal morphology can intervene between the verb's causative affix and the verb root. For instance stative morpheme *-ek* can intervene:

(13.4)
Naa-tem-*ek*-eshya iciimuti
I.PAST-cut-STAT-CAUSE stick
'I caused the stick to be cut'

[9] Actually it's not clear to me why the adverb wouldn't be able to modify *both* actions in such languages, given that they're both introduced by vPs. Pylkkänen doesn't address this.

Pylkkänen proposes an interesting connection between verb morphology and scoping possibilities in causatives: namely that in languages where a causative morpheme in the verb can be seen to attach to something that's already a verb rather than directly to a root, non-agentive adverbs will be interpreted as modifying the caused action in causatives rather than the causing action.

13.0.6.0.13 Causatives taking external-argument-introducing heads as arguments Pylkkänen's final type of language is where causatives appear even higher, above a projection introducing an external argument or a high applicative (see Section 10.1.4). Venda is an example of a language where this happens. In Section 10.1.4 we saw that high applicatives are realised morphologically on the verb. Causatives are also realised morphologically on the verb; importantly, the causative morphology is *higher* than the high applicative morphology, so we assume that the causative is higher than the applicative in phrase structure.[10]

(13.5)

tshimbila	'walk'	root
tshimbi-dz-a	'make walk'	root-CAUSE
tshimbil-el-a	'walk for'	root-APPL
tshimbil-e-dz-a	'make walk for'	root-APPL-CAUSE

(Note: we know the applicative is high, because it applies to the unergative verb *walk*.) Pylkkänen's prediction is that in a language like this, agentive adverbs like *willingly* (that modify things attached *above vP*) should be able to modify the caused action in a causative construction. Recall that that was not possible in Bemba, but in Venda it is possible:

(13.6)

Muuhambadzi	o-reng-is-a	Katonga	modoro nga dzangalelo
salesman	3SG.PAST-buy-CAUSE-FV	Katonga	car with enthusiasm

(i) 'The salesman *eagerly made* Katonga buy the car eagerly'
(ii) 'The salesman made Katonga *buy the car eagerly*'

The structure of (i) is shown in 13.7: the agentive adverbial modifies the causing action. The structure of (ii) is shown in 13.8: the agentive adverbial modifies the caused action.

(13.7) [$_{vCauseP}$ salesman [$_{voiceP}$ Katonga [$_{vp}$ buy the car]] eagerly]

(13.8) [$_{vCauseP}$ salesman [$_{voiceP}$ Katonga [$_{vp}$ buy the car] eagerly]]

This latter option is not available in Bemba, because the Bemba causative appears underneath voiceP.

[10] This relies on a model in which verb inflections attach one by one as the verb raises to higher head positions (see e.g. Belletti, 1990) so that inflections further from the root belong to projections higher in phrase structure.

Chapter 14

An account of predication

14.1 The role of WM individuals in predication

I want to think about the SM process described by the sentence *The dog is brown.* I assume it starts off with attention to an individual dog in the actual world, in the usual way, involving establishment of a salient location, a classification scale (denoting singular), and activation of a property complex, from which the dominant property assembly (the type 'dog') is identified. I assume these operations are stored in a WM individual in the normal way.

After the WM indivividual is established, it somehow participates in a SM process involving property-level IOR. The curious thing is that this process can be reported fully within the DP system (resulting in a DP containing an adjectival modifier, like *The brown dog*), or it can be reported within a full clause (resulting in a predicative clause like *The dog was brown*).[1] It's well known that there are parallels between DP structure and clause structure (see Abney, 1996 and much subsequent work). I want to think about these parallels in the context of predicative constructions.

One idea is that there's an explicit *operation*, that sets up a WM episode to record the steps in a property-level IOR operation. In the current WM model, there are operations that copy the current WM individual into a position (either agent or patient) within the WM episode medium. What might this operation be? To think about this, I'll first think about how things work in perception of an episode involving an action (i.e. something that's aspectually an event), and then think about the case of predication.

14.1.1 Perceiving a reach-to-grasp action

When the observer is experiencing a reach-to-grasp action, he first attends to the agent (setting up a WM individual, and copying it to the agent part of the WM episode medium), then to the patient (loading a new representation into the WM individual medium, and

[1] I think a similar situation arises for existentials: a new individual can be reported fully within a DP (e.g. *a dog in the garden*) or within a whole clasue (e.g. the existential *There was a dog in the garden*).

copying *this* into the patient part of the WM episode medium), then activating an action. An interesting question: at what point does the observer realise that an episode is taking place? I think it's at the very start. For instance, when the observer decides to do an action himself, and makes himself the agent, he's already decided that an episode is going to occur. But even when the observer attends to an external agent, he does so because this agent is the most salient element of an event that's already taking place. In some sense, episode perception drives the whole SM process described in *The man grabbed a cup.*[2]

14.1.2 Perceiving a property

However, when the observer perceives that a property is possessed by an individual, the episode perception process is *not* in the driving seat. The observer is just *looking at an object*: his active perceptual process is driven by his *interest* in the object, rather than by any event in which this object happens to be participating. This is really very different, and we can expect the structure of the perceptual process to be different accordingly.

My idea is that having established a WM individual, the observer can execute a special operation that sets up the WM episode medium to record a property-level IOR action and its consequences. Ultimately I'm aiming to explain why in sentences featuring the copula, the copula is inflected to agree with the subject—as is the property expression in many languages (e.g. the adjective in Romance languages, or the predicate nominal in almost all languages).

14.1.3 Aside: syntactic agreement and semantic agreement

As regards gender, there's an important difference between conceptual gender and grammatical gender. In English, gender reflects semantic properties (e.g. *he* and *she* describe a male female person respectively), while in other languages, gender is partly arbitrary. For instance in Italian, *la donna* ('the woman') is feminine because the word refers to a semantically female individual, but *la macchina* ('the car') is also feminine, despite having no semantic associations with femininity; in this case, gender is a purely grammatical attribute.

A similar distinction can be made (more rarely) for number. For instance, in French there are some nominals that are grammatically plural but semantically singular—for instance the polite form of the singular second-person pronoun (*vous*) is syntactically plural.

Vigliocco and Franck (1999) consider gender agreement in predicate nominals in French and Italian, in a sentence generation task. They find subjects make fewer agreement errors in cases where the subject has a semantic gender that aligns with its syntactic gender,

[2] An interesting intermediate case is when the observer looks at an object, *waiting for something to happen*, i.e. waiting to perceive an episode in which this object is a participant. I like to think this corresponds to a case with particular information structure. Viz: *A wolf came out of the forest. The wolf roared a terrible roar*. This is still an episode, and the observer's processing is still driven by an event happening in the world, but the observer in this case knows to *wait* for this event.

suggesting that the agreement mechanism accesses semantic information as well as just the syntactic attributes of words.

Several theorists within the HPSG tradition model semantic agreement and syntactic agreement as separate mechanisms. Number agreement in predicate nominals is an interesting case, considered by Van Eynde (2012). Van Eynde considers the number agreement in *His brothers*[PL] *are*[PL] *engineers*[PL] as semantic. In fact it's not really agreement; the plurality of the predicate nominal just reflects the fact that the property semantically plural; in a case like *The hooligans*[PL]*] are [a problem]*[SING] there is no number agreement. (It can work the other way round too, in a case like *[The French team]*[SING] *was [three farmers from the neighbouring village]*[PL]. In a case like the French *Vous[*PL*] êtes*[PL] *heureux*[SING], semantic factors dominate morphosyntactic agreement.

it isn't always—as is nicely pointed out by

Chapter 15

Tense

15.1 Simple present and simple past

15.2 The composed past in French

In the composed past, there's an **auxiliary** (*avoir* or *être*) that introduces a **past participle**: an open-class verb head whose inflections are somewhat different from normal. The past participle is always inflected for past (with the inflection -*é*. It can also carry an additional inflection agreeing with the number and gender of the subject (if the auxiliary is *être*), or with the number and gender of the object (if the auxiliary is *avoir*). The cases in which this latter agreement is present are very interesting: it only appears when the object moves to a position outside the VP—or at least, is pronounced at a higher position.

	Marie	a	fait/*-e	la robe.
(15.1)	Marie	has	made/*-FEM	the dress[FEM]
	'Marie made the dress'			

	La robe	a	été	fait-e.
(15.2)	The dress[FEM]	has	been	made-FEM
	'The dress has been made'			

	Marie	l'a	fait-e.
(15.3)	Marie	it-has	made-FEM
	'Marie made it'		

	Combien de tables	Paul	a	repeint-es?
(15.4)	How many of tables[FEM,PLUR]	Paul	has	repainted-FEM,PLUR
	'How many tables has Paul repainted?'			

I haven't considered passives yet, or wh-movement, so I'll focus on Example **??**, where the object appears as a clitic on the auxiliary verb rather than as a complement of the open-class verb.

An account of past participle agreement has to explain why these participles agree sometimes with the subject and sometimes with the object—and in the latter case, why agreement only happens when the object raises to, or is pronounced in, a position outside VP. There are a great many accounts, many of them originating in the elegant model of Kayne (1989). I want to preserve some features of this account, but also make use of an account of surface structure…

Here's my proposal about what the LF of a passé composé clause looks like. The X-bar structure is no different to that of a simple verb: re's TP, AgrSP, AgrOP, and VP. TP signals establishment of episodic memory mode; AgrSP and AgrOP signal attention to agent and patient respectively; VP signals monitoring of the motor routine. At PF, if there's a full object DP, as in *Marie a fait la robe*:

- The subject DP is pronounced at [Spec,AgrSP], and the auxiliary at AgrS (along with subject agreement inflections)
- The past participle is pronounced at V (the lower subject position where full inflected verbs are pronounced in English)
- The object DP is pronounced at the complement of V (the low object position where it's normally pronounced).

If the object is cliticised, as in *Marie l'a fait-e*:

- The subject DP is pronounced at [Spec,AgrSP]. The auxiliary (plus subject agreement inflections) is pronounced at AgrS. In addition, the clitic object pronoun is produced at this position (before the auxiliary is produced).
- The past participle is pronounced at V, as before, but this time the agreement features associated with the object are pronounced.

Linguists work hard to explain why object agreement features are only pronounced on the past participle when the object raises to a higher location. But in my model there's no actual raising process, and therefore these accounts are not available. In the model I propose, there are no strong constraints on when agreement features are read out: it's a matter of learned convention for any given language. The only key thing is that the convention be defined in a way that generalises over lexical items. (You don't have to learn it for each word.) There are various ways this could be done. One is to have a separate entropy measure for inflections, so that choosing what inflection to pronounce is fully separated from choosing what verb stem to pronounce. Another is to learn operations that either pronounce or withold various different inflections, as well as learning the inflections themselves.

Semantically, I want to argue that the passé composé is no different to the simple past: it describes events that are aspectually imperfective, just like the simple past. (The English composed past is interpreted differently, as I will discuss in Section 15.3.) I want to interpret the past auxiliary as

Question: why do some verbs get introduced by *avoir* and some by *être*? I don't think I want to link this distinction too strongly to the unaccusativity of verbs; see Levin and Rappaport-Hovav (1995) for a discussion about this issue. But an explanation of some kind is still needed.

15.3 The present perfect in English

The English present perfect is syntactically quite similar to the French passé composé, in that it involves an auxiliary verb (always *have*, at least in modern English) that introduces a past participle. But aspectually, as already mentioned, it is very different: a present perfect sentence describes an event that happened in the past whose effects continue to have relevance at the time of speech (see classically Reichenbach, 1947, and Steedman, 2005 for a good recent account).

I really want to make reference to the WM situation in my account of the present perfect. The notion that a completed episode 'continues to have relevance' at the current time, for a current agent, means that it *affects his cognitive set*. This is something that is (or should be) very well represented in my model of WM situations: it's something that alters the distribution of candidate episodes in the candidate WM episodes buffer. One idea is that there is a whole medium representing 'completed episodes', that parallels the candidate WM individuals buffer, that supports the implementation of complex plans. (This is something I talked about in Part 2.)

At the same time, the use of the auxiliary *have* in the present perfect is also important. 'Have' denotes possession, which in my model is represented through the device of indexing: if X has Y, then X (as an individual) is linked in the LTM environments system to Y (as an object). I like the idea that an individual can possess (at the reference time) not only regular physical object and sub-parts, but also episodes in which it has participated, whose current relevance is ongoing. But how this notion of possession relates to the WM situations model is still not clear. One possibility is that the completed episodes relevant to individual I are only activated in the WM situation medium when I is considered, because they're relevant to interaction with I—so this is why they're indexed to I.

Chapter 16

Machinery supporting an account of relative clauses

The memory model must be able to support generation of a sentence containing a relative clause—for instance *The man who loves the woman loves cheese.* Our key idea is that we don't have to represent all the semantic components of this sentence *explicitly* at the same time. Specifically, we suggest that we represent the content of the matrix clause and of the subordinate clause explicitly at different times. In this chapter we describe how we envisage doing this.

16.1 Semantic LTM, and episode-level properties

We think of semantic LTM as long-term memory for the properties of individuals. It's implemented as associations between LTM individuals and properties. This supports selection of the most interesting property of a LTM individual.

There are simple properties, like 'big' or 'happy'. But we also envisage that an individual's participation in an episode can be registered as a property of this individual—if it's sufficiently unusual for an individual of this type to participate in an episode of this type. To do this, we envisage a bank of connections from LTM individuals to the candidate WM episodes buffer. The episodes linked to by these connections are all **abstracted WM episodes**, in which exactly one participant is a reference (perhaps pronominal) to the individual whose property is being expressed.

16.1.1 Formation of abstracted WM episodes

There must be a mechanism that registers when an individual's participation in an episode is unusual enough to warrant creation of an episode-level property. Not every episode warrants it. For instance, I often pick up cups, but this doesn't distinguish me from other men.

Martin has an idea about how unusualness of an individual's participation in an episode can be registered. The mechanism involves inhibition of the episodes normally associated with this type of individual, or something like that. The mechanism turns on the way the WM episodes medium is linked separately to SOMs representing agent, patient and action, allowing generalisations over episodes to be expressed. (More on this from Martin.)

16.1.2 References to the individual associated with an episode-level property

Exactly one of the participants in an episode-level property is abstracted away from, as already mentioned. But how this abstraction is represented must still be decided. I like the idea that the abstracted participant has the same semantics as a prounoun. (This would explain how it can surface in some languages as a resumptive pronoun: e.g. *The man who *he loves the woman.*) Provided no ordinary references to pronouns can be encoded in episode-level properties, this should work.

But it's still important to guarantee that exactly one participant is represented as a pronoun. I'm not sure how this could be done. I think it can perhaps be guaranteed by the fact that one of the participants is going to be identical to the LTM individual associated with it; we just have to make sure that other participants are different. (E.g. in *A man who [X loves himself]* the reflexive property of the object is crucial.)

16.1.3 Activating an episode-level property

There's something you can do to see if there are any interesting episode-level properties. You activate a LTM individual, activate its associated distribution of episode-level properties, pick the winner, and see if it exceeds some threshold.

16.1.4 Implicit episode-level properties

The representation of an LTM individual doesn't make explicit any of its episode-level properties, but they're *implicitly* represented, in its connections with the candidate WM episodes buffer.

16.2 The episodic buffer (a revised model)

Let's say I experience an episode. When I do this, I use the candidate WM episodes buffer to help generate top-down expectations—but ultimately I end up reconciling these with bottom-up experiences. At the end of this process I've got a representation of a single episode in what I'll call the **episodic buffer**, which has the capacity to hold exactly one episode. It represents all the semantic components of the episode, in distinct media: i.e. it uses binding-by-space to represent the agent and the patient (and represents the action separately as well).

16.3 Generating a sentence with a relative clause

Say we've got an experienced episode in the episodic buffer: 'The man loves cheese'. This involves a representation of the man in the 'agent' medium, cheese in the 'patient' medium, and 'loves' in the 'action' medium. In fact what we're representing is not any old man: it's a particular man, who has particular properties. These properties are not *explicitly* represented in our scheme: crucially, we can only represent the properties of one individual at a time, and there are two individuals in the episode. But each individual is represented as a unit which has links to its properties. We envisage an operation that **enables** a particular medium: when a medium is enabled, then the active unit's connections are gated open. When the agent medium is enabled, we activate the properties of the individual represented in that medium; likewise for the patient medium. We envisage the agent and patient media being activated sequentially, in turn, when the WM episode is 'rehearsed'.

When we enable the agent medium, we evoke the properties of the agent of the episode. Importantly, these include episode-level properties in the candidate WM episodes buffer. Up to now, the candidate WM episodes buffer served to help establish the episode that has just been experienced: 'The man loves cheese'. But now that the episode is represented as the winning episode (in the episodic buffer), the candidate WM episodes buffer is able to hold a completely different representation: namely the episode-level properties of the individual in the currently enabled medium. If one of these is particularly interesting, for some reason, it can be selected, and rehearsed in its own right. We suggest the decision to do this is expressed linguistically in a relative pronoun (e.g. *who*).

16.3.1 A stack of selected episodes

To rehearse an episode-level property, we have to store the matrix episode. We envisage a stack of episodes, implemented with a recursive autoassociative memory (RAAM: Pollack, 1990). When we have rehearsed the property in a relative clause, we pop the stack and continue with the matrix clause.

The key idea is that we only ever represent a single clause at a time *as a pattern of activation*.[1] In the matrix clause, the units representing the agent individual are *linked* to units which can retrieve the content of the relative clause if this is necessary, but in this case the content of the relative clause will temporarily replace that of the matrix clause.

The nice thing about this scheme is that while we have a stack of *winning episodes*, we don't need to envisage a stack of complete *episode distributions*, which would be too expensive to envisage.

[1] And in fact, even within a single clause we only represent one individual at a time as a *rich* pattern of activation; the units active in parallel in the agent and patient media can be read out linguistically as agreement inflections, and possibly as clitic pronouns if they're pronominal.

16.4 Pragmatics of relative clause generation

I'm just considering referential relative clauses so far. These are generated in order to distinguish the individual to be referred to from distractors. The representation of a sentence with or without the relative clause is the same. What differs is a representation of the pragmatic context in which the sentence is generated. If there are other salient men, then a relative clause will be produced, that identifies properties of the target man that distinguish it from the distractor men; if not, then it's not necessary to generate a relative clause.

In fact, the set of distractors depends not just on context, but on the supposed knowledge of the interlocutor. As discussed in Chapter 1, the set of WM episodes that defines a communicative sitation depends on the selected interlocutor; we envisage this principle applying in the realm of distinguishing descriptions as well as in the case of selecting a suitable matrix proposition to communicate. How this happens is still an open question.

16.5 Connections between individual-denoting media and the candidate WM episodes buffer

One important feature of the model just outlined is that there are two completely different types of connection between representations of individuals and the WM episodes buffer. One bank of connections maps an LTM individual onto a distribution of episodes in which this individual memorably participated. (From this distribution a winning 'memorable episode' can be picked.) The other bank of connections links a given unit in the candidate WM episodes buffer (representing a single episode, or episode type) onto the individuals that participate in this episode. These connections specify the *content* of the episode being represented. In itself, a unit in the candidate episodes buffer is nothing but a pointer (a **convergence zone**, in the parlance of Damasio and Damasio, 1994) to the different elements of its content. It's important that these two banks of connections are kept completely separate.[2]

16.6 SOMs for representing the dominant property assembly

In the medium representing an individual's properties, we envisage a **property complex**—a distribution of all of the properties of an individual—and a **winning property assembly**—a competitive medium in which properties that co-occur are represented together, but compete against one another. The former medium is assumed to represent token individuals; the latter represents types (both the denotations of common nouns, like *dog*, and the

[2] In fact these latter connections are subdivided into further groups, because the connections to the agent medium are quite distinct from the connections to the patient medium.

denotations of adjectives, like *furry* or *happy*). I envisage an operation of property-level IOR, that inhibits the winning property assembly, and activates in its place a new winner, encoding the properties in which the token individual being represented differs most from the prototypical item of its class.

Martin's suggestion is that we implement the winning property assembly as a SOM. Technically, this promises to be a good implementation of property-level IOR. But in addition to this, it also has two other benefits.

Firstly, it has some of the right qualities for the above account of relative clauses. When we represent the agent 'man', we're not representing any old man—it's a man with certain properties, as mentioned before. A SOM is a good means to represent these.

Secondly, it also allows us to represent generic individuals (e.g. an animal, rather than just a dog or a cat). Not sure how this one pans out yet...

16.7 Questions

16.7.1 Which medium supports property-level IOR?

I've been accustomed to thinking about the medium that supports property-level IOR as being the WM individuals medium. In my previous thinking, the property-level IOR mechanism supports the expression of predications like *The dog is hairy* or *The dog is a dachshund*. But now I'm using a similar notion to support an account of relative clauses. And the medium in question is the LTM individual. (It's the LTM indivdual that's linked to episode-level properties; I think it's important it should be that way, because individuals with similar properties as objects are still distinct token individuals, and can participate in completely different episodes.)

Of course it's possible that relative clauses are semantically different from predicative sentences. Maybe relative clauses require activation of the properties of LTM individuals, while predicative clauses don't. Syntactically they seem different (the copula in one case, a relative pronoun in the other). But syntacticians sometimes say that adjectives inside a DP originate in predicative relative clauses. This connection needs thinking about...

Chapter 17

Questions and information structure

17.1 An idea about the roles of PFC and hippocampus in episodic memory access

When we use a cue stimulus to recall material from episodic memory, we have to keep the cue stimulus separate from the retrieved material, so that we can check whether the retrieved material really is a response to the cue. This is called 'source monitoring' (see e.g. ??). PFC is given a role in source monitoring, rather than retrieval *per se*. Source monitoring suggests the cue is kept in two locations. One location holds the 'active' cue, which actually functions as a retrieval cue; in this medium, presumably, the retrieved material completes, or even overwrites, the original cue, so in this medium the cue is (potentially) changed by the retrieval mechanism. The other is the location holding a 'reference copy' of the cue, that isn't changed, and is used to compare with the retrieved material during source monitoring. If PFC holds the reference copy of the cue, and the hippocampus holds the active cue, we expect there to be a copying operation linking the two areas. Let's say language processing creates a semantic episode representation in PFC, but there's also another, non-linguistic route for creating semantic representations in WM, directly in the hippocampus. This non-linguistic route means the PFC is not essential for querying episodic memory—but for memory cues expressed in sentences, a WM representation *will* be created in PFC, and then copied to hippocampus. And this provides exactly the format needed to model focus-structures in the answers to questions.

On this model, when you perceive an episode through SM experience, it's stored independently in WM in PFC and in the hippocampus. I presume the direct link between PFC and hippocampus is active, and used to align the two WM representations, to make sure they agree. The PFC one is used in sentence generation; the hippocampal one is used to store the episode in episodic memory.

Bibliography

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.

Adams, M., Johnson, S., Lefèvre, P., Lévesque, V., Hayward, A., and Thonnard, J.-L. (2013). Finger pad friction and its role in grip and touch. *Journal of the Royal Society Interface*, **10**, 20120467.

Amedia, A., Raz, N., Azulay, H., Malach, R., and Zohary, E. (2010). Cortical activity during tactile exploration of objects in blind and sighted humans. *Restorative Neurology and Neuroscience*, **28**, 143–156.

Andersen, R. and Buneo, C. (2002). Intentional maps in posterior parietal cortex. *Annual Review of Neuroscience*, **25**, 189–220.

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.

Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J., Zecchina, R., and Zoccolan, D. (2013). Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS*, **9**(8), e1003167.

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.

Batista, A., Santhanam, G., Yu, B., Ryu, S., Afshar, A., and Shenoy, K. (2007). Reference frames for reach planning in macaque dorsal premotor cortex. *Journal of Neurophysiology*, **98**, 966–983.

Baylis, G. and Driver, J. (1993). Visual attention and objects: Evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Perception and Performance*, **19**, 451–470.

Beck, S. and Johnson, K. (2004). Double objects again. *Linguistic Inquiry*, **35**(1), 97–124.

Behrmann, M. and Tipper, S. (1999). Attention accesses multiple reference frames: Evidence from visual neglect. *Journal of Experimental Psychology*, **25**(1), 83–101.

Belletti, A. (1990). *Generalized Verb Movement*. Rosenberg and Sellier, Turin.

Bertamini, M., Helmy, M., and Bates, D. (2013). The visual system prioritizes locations near corners of surfaces (not just locations near a corner). *Attention Perception and Psychophysics*, **75**(8), 1748–1760.

Biederman, I. (1985). Human image understanding: Recent research and a theory. *Computer Vision, Graphics and Image Processing*, **32**, 29–73.

Bolognini, N., Rossetti, A., Maravita, A., and Miniussi, C. (2011). Seeing touch in the somatosensory cortex: A tms study of the visual perception of touch. *Human Brain Mapping*, **32**(12), 2104–2114.

Boneh, N. and Nash, L. (2009). A higher-order applicative: evidence from French. In *Proceedings of IATL 25*, pages 1–21.

Bouchard, D. (1995). *The Semantics of Syntax: A Minimalist Approach to Grammar*. University of Chicago Press, Chicago and London.

Bouchard, S., Talbot, J., Ledoux, A., Phillips, J., Cantames, M., and Robillard, G. (2009). The meaning of being there is related to a specific activation in the brain located in the parahypocampus. In *Proceedings of the 12th Annual International Workshop on Presence, Los Angeles (CA)*.

Brass, M. and Haggard, P. (2010). The hidden side of intentional action: the role of the anterior insular cortex. *Brain Structure and Function*, **214**, 603–610.

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.

Brooke, J. (2008). Light np shift and verbal adjuncts in english. In *Proceedings of the Conference of the Canadian Linguistics Association*.

Brozzoli, C., Gentile, G., and Ehrsson, H. (2012). That's near my hand! parietal and premotor coding of hand-centered space contributes to localization and self-attribution of the hand. *Journal of Neuroscience*, **32**(42), 14573–14582.

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**, 400–404.

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.

Carlson, G. (1977). *Reference to Kinds in English*. Ph.D. thesis, University of Massachusetts at Amherst.

Cathcart, M. (2011). *Impulsatives: The Syntax and Semantics of Involuntary Desire*. Ph.D. thesis, University of Delaware, Department of Linguisticso.

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.

Chen, Z. (2012). Object-based attention: A tutorial review. *Attention, Perception and Psychophysics*, **74**, 784–802.

Chen, Z., Gomperts, S., Yamamoto, J., and Wilson, M. (2014). Neural representation of spatial topology in the rodent hippocampus. *Neural Computation*, **26**, 1–39.

Ciaramitaro, V., Mitchell, J., Stoner, G., Reynolds, J., and Boynton, J. (2011). Object-based attention to one of two superimposed surfaces alters responses in human early visual cortex. *Journal of Neurophysiology*, **105**(3), 1258–1265.

Cinque, G. (1994). 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.

Cinque, G. and Rizzi, L., editors (2010). *Mapping Spatial PPs: The Cartography of Syntactic Structures Vol. 6*. Oxford University Press, Oxford, UK.

Cohen, D., Prud'Homme, M., and Kalasaka, J. (1994). Tactile activity in primate primary somatosensory cortex during active arm movements: Correlation with receptive field properties. *Journal of Neurophysiology*, **71**(1), 161–172.

Colby, C. and Goldberg, M. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience*, **22**, 319–349.

Corbetta, M. and Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, **3**, 201–215.

Cosman, J. and Vecera, S. (2010). Attention affects visual perceptual processing near the hand. *Psychological Science*, **21**(9), 12541258.

Craig, A. (2009). How do you feel—now? the anterior insula and human awareness. *Nature Reviews Neuroscience*, **10**, 59–70.

Critchley, H., Wiens, S., Rotshtein, P., Öhman, A., and Dolan, R. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, **7**(2), 189–195.

Dabaghian, Y., Brandt, V., and Frank, L. (2014). Reconceiving the hippocampal map as a topological template. *eLife*, **3**, e03476.

Damasio, A., editor (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace, New York.

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*, pages 61–74. MIT Press, Cambridge, MA.

Deliagina, T., Orlovsky, G., Zelenin, P., and Beloozerova, I. (2006). Neural bases of postural control. *Physiology*, **21**, 216–225.

Demeyer, M., De Graef, P., Verfaillie, K., and Wagemans, J. (2011). Perceptual grouping of object contours survives saccades. *PLoS One*, **6**(6), e21257.

den Dikken, M. (2003). On the syntax of locative and directional adpositional phrases. Ms., CUNY Graduate Center.

den Dikken, M. (2010). On the functional structure of locative and directional PPs. In G. Cinque and L. Rizzi, editors, *Mapping Spatial PPs: The Cartography of Syntactic Structures Vol. 6*, pages 74–126. Oxford University Press, Oxford, UK.

Deneve, S. and Pouget, A. (1992). Basis functions for object-centered representations. *Neuron*, **37**, 347–359.

Derdikman, D. and Moser, E. (2010). A manifold of spatial maps in the brain. *Trends in Cognitive Sciences*, **14**(12), 561–569.

Deubel, H., Schneider, W., and Paprotta, I. (1998). Evidence for a common attentional mechanism in reaching and perception. *Visual Cognition*, **5**, 81–107.

Diba, K. and Buzsàki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, **10**(10), 1241–1242.

DiCarlo, J., Zoccolan, D., and Rust, N. (2012). How does the brain solve visual object recognition? *Neuron*, **73**, 415–434.

Doeller, C. and Kaplan, R. (2011). Parahippocampal cortex: translating vision into space. *Current Biology*, **21**(15), R590.

Dowty, D. (1979). *Word Meaning and Montague Grammar*. Reidel, Dordrecht.

Driver, J., Baylis, G., Goodrich, S., and Rafal, R. (1994). Axis-based neglect of visual shapes. *Neuropsychologia*, **32**(11), 1353–1365.

Dudchenko, P. and Zinyuk, L. (2005). The formation of cognitive maps of adjacent environments: Evidence from the head direction cell system. *Behavioral Neuroscience*, **119**(6), 1511–1523.

Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. *Nature Reviews Neuroscience*, **15**, 732–744.

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.

Fagg, A. and Arbib, M. (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Networks*, **11**(7/8), 1277–1303.

Fenton, A., Lytton, W., Barry, J., Lenck-Santini, P.-P., Zinyuk, L., Kubik, S., Bures, J., Poucet, B., Muller, R., and Olypher, A. (2010). Attention-like modulation of hippocampus place cell discharge. *Journal of Neuroscience*, **30**(13), 4613–4625.

Ferbinteanu, J. and Shapiro, M. (2003). Prospective and retrospective memory coding in the hippocampus. *Neuron*, **40**, 1227–1239.

Fitzgerald, P., Lane, J., Thakur, P., and Hsiao, S. (2006). Receptive field (rf) properties of the macaque second somatosensory cortex: Rf size, shape, and somatotopic organization. *Journal of Neuroscience*, **26**(24), 6485–6495.

Fodor, J. (1970). Three reasons for not deriving ‘kill’ from ‘cause to die’. *Linguistic Inquiry*, **1**, 429–438.

Foster, D. and Wilson, M. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, **440**, 680–683.

Frank, L., Brown, E., and Wilson, M. (2000). Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron*, **27**(1), 169–178.

Fyhn, M., Molden, S., Witter, M., Moser, E., and Moser, M. (2004). Spatial representation in the entorhinal cortex. *Science*, **305**, 1258–1264.

Gehrke, B. (2008). *Ps in Motion: on the semantics and syntax of P elements and motion events*. Ph.D. thesis, University of Utrecht.

Gibson, J., editor (1950). *The perception of the visual world*. Houghton Mifflin, Boston.

Gillmeister, H. and Forster, B. (2012). Adverse effects of viewing the hand on tactile-spatial selection between fingers depend on finger posture. *Experimental Brain Research*, **221**, 269–278.

Givón, T. (1976). Some constraints on Bantu causativization. In M. Shibatani, editor, *Syntax and Semantics Vol. 6: Grammar of Causative Constructions*. Academic Press, New York.

Gopnik, A. and Wellman, H. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms and the theory theory. *Psychological Bulletin*, **138**(6), 1085–1108.

Gottlieb, J., Kusunoki, M., and Goldberg, M. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, **391**, 481–484.

Graziano, M., Hu, X., and Gross, C. (1997). Visuospatial properties of ventral premotor cortex. *Journal of Neurophysiology*, **77**, 2268–2292.

Haggard, P. (2006). Sensory neuroscience: From skin to object in the somatosensory cortex. *Current Biology*, **16**(20), R884R886.

Haggard, P., Kitadono, K., Press, C., and Taylor-Clarke, M. (2006). The brains fingers and hands. *Experimental Brain Research*, **172**, 94–102.

Haggard, P., Christakou, A., and Serino, A. (2007). Viewing the body modulates tactile receptive fields. *Experimental Brain Research*, **180**, 187–193.

Hamada, Y. and Suzuki, R. (2005). Hand posture modulates cortical finger representation in SII. *Neuroimage*, **25**, 708–717.

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.

Hartley, T., Maguire, E., Spiers, H., and Burgess, N. (2003). The well-worn route and the path less traveled: Distinct neural bases of route following and wayfinding in humans. *Neuron*, **37**, 877–888.

Hassabis, D., Kumaran, D., Vann, S., and Maguire, E. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of the USA*, **104**(5), 1726–1731.

He, Z. and Nakayama, K. (1995). Visual attention to surfaces in three-dimensional space. *Proceedings of the National Academy of Sciences of the USA*, **92**, 11155–11159.

Henderson, J. and Ferreira, F. (2004). Scene perception for psycholinguists. In J. Henderson and F. Ferreira, editors, *The Interface of Language, Vision and Action: Eye Movements and the Visual World*, pages 1–58. Psychology Press, New York.

Hirtle, S. and Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory and Cognition*, **13**(3), 208–217.

Hume, D. (1777/1975). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Clarendon Press, Oxford.

Hung, C.-C., Carlson, E., and Connor, C. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, **74**, 1099–1113.

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.

Iriki, A., Tanaka, M., and Iwamura, Y. (1996). Coding of modified body schema during tool use by macaque postcentral neurones. *Neuroreport*, **7**(14), 2325–2330.

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.

Iwamura, Y., Taoka, M., and Iriki, A. (2001). Bilateral activity and callosal connections in the somatosensory cortex. *Neuroscientist*, **7**(5), 419–429.

J, K., Ghim, J.-W., Lee, J., and Jung, M. (2013). Neural correlates of interval timing in the prefrontal cortex. *Journal of Neuroscience*, **33**, 13834–13847.

Jackendoff, R. (1973). The base rules for prepositional phrases. In S. Anderson and P. Kiparsky, editors, *A festschrift for Morris Hallé*, pages 345–356. Holt, Rinehart and Winston, New York.

Jackendoff, R. (1983). *Semantics and cognition*. MIT Press, Cambridge, MA.

Jackendoff, R. (1990). *Semantic Structures*. MIT Press, Cambridge, MA.

Jeffery, K. (2011). Place cells, grid cells, attractors, and remapping. *Neural Plasticity*, **2011**, 182602.

Jeffery, K., Jovalekic, A., Verriotis, M., and Hayman, R. (2013). Navigating in a three-dimensional world. *Behavioural and Brain Sciences*, **36**, 523–587.

Jehee, J., Lamme, V., and Roelfsema, P. (2007). Boundary assignment in a recurrent network architecture. *Vision Research*, **47**, 1153–1165.

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.

Ji, D. and Wilson, M. (2008). Firing rate dynamics in the hippocampus induced by trajectory learning. *Journal of Neuroscience*, **21**(18), 4679–4689.

Jordan, M. and Wolpert, D. (2000). Computational motor control. In M. Gazzaniga, editor, *The New Cognitive Neurosciences*, pages 71–118. MIT Press, Cambridge, MA.

Kayne, R. (1989). Facets of past participle agreement in Romance. In P. Benincà, editor, *Dialect variation and the theory of grammar*. Foris, Dordrecht.

Kimchi, R. and Peterson, M. (2008). Figure-ground segmentation can occur without attention. *Psychological Science*, **19**(7), 660–668.

Kiparsky, P. (1998). Partitive case and aspect. In M. Butt and W. Geuder, editors, *The Projection of Arguments*. CSLI, Stanford.

Kleinke, C. and Pohlen, P. (1971). Affective and emotional responses as a function of other persons gaze and cooperativeness in a two-person game. *Journal of Personality and Social Psychology*, **17**, 308–313.

Knott, A. (2012). *Sensorimotor Cognition and Natural Language Syntax*. MIT Press, Cambridge, MA.

Koopman, H. (2000). Prepositions, postpositions, circumpositions, and particles. In H. Koopman, editor, *The Syntax of Specifiers and Heads*, pages 204–260. Routledge, London, New York.

Kratzer, A. (1996). Severing the external argument from its verb. In J. Rooryck and L. Zaring, editors, *Phrase structure and the lexicon*, pages 109–137. Kluwer Academic Publishers.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In *Proceedings of ICCV 2013, Workshop on 3D Representation and Recognition*.

Krifka, M. (1992). Thematic relations as links between nominal reference and temporal constitution. In I. Sag and A. Szabolcsi, editors, *Lexical matters*, pages 29–53. CSLI Publications, University of Chicago Press.

Lacey, S andTal, N., Amedi, A., and Sathian, K. (2009). A putative model of multisensory object representation. *Brain Topography*, **21**, 269–274.

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.

Landau, I. (2009). *The Locative Syntax of Experiencers*. MIT Press, Cambridge, MA.

Larson, R. (2010). On Pylkkänen's semantics for low applicatives. *Linguistic Inquiry*, **41**, 701–704.

Lederman, S. and Klatzky, R. (1993). Extracting object properties through haptic exploration. *Acta Psychologica*, **84**(1), 29–40.

Lederman, S. and Klatzky, R. (2011). Haptic perception: A tutorial. *Attention, Perception and Psychophysics*, **71**(7), 1439–1459.

Lee-Hand, J. and Knott, A. (2013). A model of causative actions: from motor learning to syntactic structure. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, pages 2849–2854, Berlin.

Lee-Hand, J. and Knott, A. (2015). A neural network model of causative actions. *Frontiers in Neurorobotics*, **9**, Article 4.

Lehky, S. and Sereno, A. (2007). A comparison of shape encoding in primate dorsal and ventral visual pathways. *Journal of Neurophysiology*, **97**, 307–319.

Levesque, H., Pirri, F., and Reiter, R. (1998). Foundations for the situation calculus. *Electronic Transactions on Artificial Intelligence*, **2**(3–4), 159–178.

Levin, B. and Rappaport Hovav, M. (1995). *Unaccusativity: At the syntax-lexical semantics interface*. MIT Press, Cambridge, MA.

Li, S., Ostwald, D., Giese, M., and Kourtzi, Z. (2007). Flexible coding for categorical decisions in the human brain. *Journal of Neuroscience*, **27**(45), 2321–12330.

Lipton, J. and Spelke, E. (2005). Preschool children's mapping of number words to non-symbolic numerosities. *Child Development*, **76**(5), 978–988.

MacDonald, C., Lepage, K., Eden, U., and Eichenbaum, H. (2011). Hippocampal 'time cells' bridge the gap in memory for discontiguous events. *Neuron*, **71**, 737–749.

Maguire, E., Burgess, N., Donnett, J., Frackowiak, R., Frith, C., and O'Keefe, J. (1998). Knowing where and getting there: A human navigation network. *Science*, **280**(5365), 921–924.

Marr, D. (1982). *Vision*. Freeman.

Martin, V., Schacter, D., Corballis, M., and Addis, D. (2011). A role for the hippocampus in encoding simulations of future events. *Proceedings of the National Academy of the USA*, **108**(33), 13858–13863.

Mattell, M., , and Meck, W. (2000). Neuropsychological mechanisms of interval timing behavior. *BioEssays*, **22**(1), 94–103.

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.

Mendez, M. and Cherrier, M. (2003). Agnosia for scenes in topographagnosia. *Neuropsychologia*, **41**, 13871395.

Menon, V. and Uddin, L. (2009). The anterior insula in autism: Under-connected and under-examined. *Neuroscience and Biobehavioral Reviews*, **33**(8), 1198–1203.

Menon, V. and Uddin, L. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, **214**, 655–667.

Moens, M. and Steedman, M. (1987). Temporal ontology in natural language. In *Proceedings of the 25th ACL Conference*, pages 1–7, Stanford, CA.

Moore, T. and Armstrong, K. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, **421**(6921), 370–373.

Muller, R. and Kubie, 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., 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.

Nadel, L. (1991). The hippocampus and space revisited. *Hippocampus*, **1**(3), 221–229.

Navalpakkam, V. and Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, **45**, 205–231.

Nelson, D. (1998). *Grammatical Case Assignment in Finnish*. Taylor and Francis, Florence, KY.

O'Craven, K., Downing, P., and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, **401**, 584–587.

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.

Orban, G. (2011). The extraction of 3d shape in the visual system of human and nonhuman primates. *Annual Review of Neuroscience*, **34**, 361–388.

Oztop, E., Bradley, N., and Arbib, M. (2004). Infant grasp learning: a computational model. *Experimental Brain Research*, **158**, 480–503.

OHerron, P. and von der Heydt, R. (2013). Remapping of border ownership in the visual cortex. *Journal of Neuroscience*, **33**(5), 1964–1974.

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.

Park, S., Brady, T., Greene, M., and Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, **31**(4), 1333–1340.

Partee, B. (1984). Nominal and temporal anaphora. *Linguistics and Philosophy*, **7**(3), 243–286.

Paz-Villagràn, V., Save, E., and Poucet, B. (2004). Independent coding of connected environments by place cells. *European Journal of Neuroscience*, **20**, 1379–1390.

Penfield, W. and Rasmussen, T. (1950). *The cerebral cortex of man; a clinical study of localization of function.*. Macmillan, New York.

Penhune, V. and Steele, C. (2012). Parallel contributions of cerebellar, striatal and m1 mechanisms to motor sequence learning. *Behavioural Brain Research*, **226**(2), 579–591.

Pereira, A., James, K., Jones, S., and Smith, L. (2010). Early biases and developmental changes in self-generated object views. *Journal of Vision*, **10**, 1–13.

Perrone, J. (1992). Model for the computation of self-motion in biological systems. *Journal of the Optical Society of America*, **9**(2), 177–194.

Peterson, M. and Kim, J. (2001). On what is bound in figures and grounds. *Visual Cognition*, **8**(3/4/5), 329–348.

Pollack, J. (1990). Recursive distributed representations. *Artificial Intelligence*, **46**(1–2), 77–105.

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press, Cambridge, MA.

Pylkkänen, L. (2002). *Introducing Arguments*. Ph.D. thesis, Dept of Linguistics and Philosophy, MIT.

Pylkkänen, L. (2008). *Introducing Arguments*. MIT Press, Cambridge, MA.

Rainer, G., Asaad, W., and Miller, E. (1998). Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, **393**, 577=–599.

Reichenbach, H. (1947). *Elements of symbolic logic*. University of California Press, Berkeley, CA.

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.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, **2**, 661–670.

Rochefort, C., Lefort, J., and Rondi-Reig, L. (2013). The cerebellum: a new key structure in the navigation system. *Frontiers in Neural Circuits*, **13**, Article 35.

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. (1997). Spatial view cells in the primate hippocampus. *European Journal of Neuroscience*, **9**, 1789–1794.

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.

Sands, K. and Campbell, L. (2001). Non-canonical subjects and objects in Finnish. In A. Aikhenvald, R. Dixon, and M. Onishi, editors, *Non-Canonical Marking of Subjects and Objects*, pages 251–305. Benjamins, Amsterdam.

Scholl, B. (2001). Objects and attention: the state of the art. *Cognition*, **80**(1–2), 1–46.

Schultz, W., Dayan, P., and Montague, P. (1997). A neural substrate of prediction and reward. *Science*, **275**, 1593–1599.

Sharp, P. (1997). Subicular place cells expand or contract their spatial firing pattern to fit the size of the environment in an open field but not in the presence of barriers: comparison with hippocampal place cells. *Behavioural Neuroscience*, **113**(4), 643–62.

Shimansky, Y., Saling, M., Wunderlich, D., Bracha, V., Stelmach, G., and Bloedel, J. (1997). Impaired capacity of cerebellar patients to perceive and learn two-dimensional shapes based on kinesthetic cues. *Learning and Memory*, **4**, 36–48.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R., and Frith, C. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, **303**(5661), 1157–1162.

Singer, T., Critchley, H., and Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences,*, **13**(8), 334–340.

Steedman, M. (2005). The productions of time. Draft tutorial notes about temporal semantics, Draft 5.0.

Sutton, R. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, **3**, 9–44.

Svenonius, P. (2007). Projections of P. In A. Asbury, J. Dotlačil, B. Gehrke, and R. Nouwen, editors, *Syntax and Semantics of Spatial P*, pages 63–84. John Benjamins, Amsterdam/Philadelphia.

Taube, J. (2007). The head direction signal: Origins and sensory-motor integration. *Annual Review of Neuroscience*, **30**, 181–207.

Taube, J., Muller, R., and Ranck, J. (1990). Head direction cells recorded from the postsubiculum in freely moving rats. 1. Description and quantitative analysis. *Journal of Neuroscience*, **10**, 420–435.

Theys, T., Srivastava, S., van Loon, J., Goffin, J., and Janssen, P. (2012). Selectivity for three-dimensional contours and surfaces in the anterior intraparietal area. *Journal of Neurophysiology*, **107**(3), 995–1008.

Thompson, K. and Bichot, N. (2005). A visual salience map in the primate frontal eye field. *Progress in Brain Research*, **147**, 251–262.

Tillery, S., Soechting, J., and Ebner, T. (1996). Somatosensory cortical activity in relation to arm posture: Nonuniform spatial tuning. *Journal of Neurophysiology*, **76**(4), 2423–2438.

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.

Tortora, C. (2006). Aspect inside PLACE PPs. Paper presented at the 'Syntax and semantics of spatial P' conference, Utrecht, 24 June 2006. Also available as ms., College of Staten Island The Graduate Center, CUNY.

Van Eynde, F. (2012). On the agreement between predicative complements and their target. In S. Müller, editor, *Proceedings of the 19th International Conference on Head-Driven Phrase Structure Grammar*, pages 349–367, Stanford. CSLI Publications.

van Riemsdijk, H. (1978). *A case study in syntactic markedness: the Binding Nature of Prepositional Phrases*. Foris Publications, Dordrecht.

van Riemsdijk, H. (1990). Functional prepositions. In H. Pinkster and I. Genée, editors, *Functional Diversity*, pages 229–241. Foris, Dordrecht.

van Riemsdijk, H. (1999). Clitics: A state of the art report. In H. van Riemsdijk, editor, *Clitics in the Languages of Europe*, pages 1–30. Mouton de Gruyter, Berlin.

van Riemsdijk, H. (2007). Case in spatial adpositional phrases: The dative-accusative alternation in German. In G. Alboiu, A. Avram, L. Avram, and D. Isac, editors, *Pitar Mos: A building with a view. Papers in honor of Alexandra Cornilescu*, pages 265–283. Bucharest University Press, Bucharest, Romania.

van Riemsdijk, H. and Huygbregts, R. (2001). Location and locality. In M. van Oostendorp and E. Anagnostopoulou, editors, *Progress in Grammar*, pages 1–23. Meertens/KNAW, Amsterdam.

Vieville, T. (2006). About biologically plausible trajectory generators. In *Proceedings of the International Joint Conference on Neural Networks*, pages 563–572, Vancouver, Canada.

Vigliocco, G. and Franck, J. (1999). When sex and syntax go hand in hand: Gender agreement in language production. *Journal of Memory and Language*, **40**, 455–478.

Waldmann, M., Cheng, P., Hagmayer, Y., and Blaisdell, A. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater and M. Oaksford, editors, *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, pages 453–484. Oxford University Press, Oxford, UK.

Westergaard, G., Liv, C., Haynie, M., and Suomi, S. (2000). A comparative study of aimed throwing by monkeys and humans. *Neuropsychologia*, **38**, 1511–1517.

Whitlock, J., Sutherland, R., Witter, M., Moser, M.-B., and Moser, E. (2008). Navigating from hippocampus to parietal cortex. *Proceedings of the National Academy of the USA*, **105**(39), 14755–14762.

Wikenheiser, A. and Redish, D. (2013). The balance of forward and backward hippocampal sequences shifts across behavioral states. *Journal of Cognitive Neuroscience*, **23**, 22–29.

Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, **14**, 715–770.

Wunderlich, D. (1991). How do prepositional phrases fit into compositional syntax and semantics? *Linguistics*, **29**, 591–621.

Yamane, Y., Carlson, E., Bowman, K., Wang, Z., and Connor, C. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, **11**(11), 1352–1360.

Yau, J., Connor, C., and Hsiao, S. (2013a). Representation of tactile curvature in macaque somatosensory area 2. *Journal of Neurophysiology*, **109**(12), 2999–3012.

Yau, J., Pasupathy, A., Brincat, S., and Connor, C. (2013b). Curvature processing dynamics in macaque area V4. *Cerebral Cortex*, **23**, 198–209.

Yoder, R., Clark, B., Brown, J., Lamia, M., Valerio, S., Shinder, M., and Taube, J. (2011). Both visual and idiothetic cues contribute to head direction cell stability during navigation along complex routes. *Journal of Neurophysiology*, **105**, 2989–3001.

Zacks, J., Kurby, C., Eisenberg, M., and Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience*, **23**(12), 4057–4066.

Zhang, Y., Meyers, E., Bichot, N., Serre, T., Poggio, T., and Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences of the USA*, **108**(21), 8850–8855.

Zhou, H., Friedman, H., and von der Heydt, R. (2001). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, **20**, 6594–6611.