

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Emulation and Cognition

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy in

Cognitive Science and Philosophy

by

Rick Grush

Committee in charge:

Professor Paul M. Churchland, Chair
Professor Patricia Smith Churchland
Professor Garrison Cottrell
Professor Adrian Cussins
Professor Robert Hecht-Nielsen
Professor Vilayanur Ramachandran

1995

Copyright ©
Rick Grush, 1995
All rights reserved.

The dissertation of Rick Grush is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

1995

Table of Contents

Signature Page	iii
Table of Contents.....	iv
List of Figures.....	vi
Acknowledgments.....	viii
Vita, Publications and Fields of Study.....	x
Abstract	xi
Chapter One: Preliminaries	1
1.1 Introduction.....	1
1.2 Brief history of this project	3
1.3 Survey of the current project.....	5
Chapter Two: Emulation and Control	11
2.1 A control thought-experiment.....	11
2.2 Inverse vs. forward mappings.....	13
2.3 What is an emulator.....	15
2.4 Benefits of emulation.....	19
2.5 Emulation and neural networks.....	23
Chapter Three: Perception, Imagery, and the Sensorimotor Loop.....	29
3.1 Motor control.....	29
3.2 Motor imagery.....	36
3.3 Visual imagery.....	39
3.4 Perception and closed-loop imagery.....	46
3.5 Conclusion.....	49
Chapter Four: Emulation, Representation and Learning	50
4.1 Dreaming, pretense, and altered control loops.....	50
4.2 Representational redescription.....	51
4.3 Ways of emulator making.....	56
4.4 What can ETM say about development?.....	62
4.5 Parameters, distributed representations and semi-locality.....	66
4.6 Concluding remarks	68

Chapter Five: The Mind-Body Solution.....	70
5.1 How the folk get their psychology.....	71
5.2 Josef Perner: Modeling models.....	72
5.3 Further thoughts on Leslie, Wellman and the simulation theory ...	78
5.4 Emulating emulators.....	81
5.5 Conclusion.....	82
 Chapter Six: The Grammar of Thought.....	 84
6.1 Innateness and syntactic autonomy	85
6.2 Anaphora and extraction.....	88
6.3 Outline of the GB account.....	93
6.4 Cognitive grammar.....	106
6.5 Binding.....	112
6.6 Temporal characteristics of language processing.....	119
6.7 What all of this buys us	122
6.8 Conclusion.....	132
 Chapter Seven: Semantics	 134
7.1 Neo-Fregean theories of meaning.....	135
7.2 X-role semantics	141
7.3 Towards a positive theory.....	146
7.4 Interpretational semantics	148
7.5 Problems with interpretational semantics.....	153
7.6 The general interpretational semantic theory (GIST).....	154
7.7 The prospects for communication.....	171
7.8 The role of the interpreter	172
7.9 Conclusion.....	174
 References.....	 176

LIST OF FIGURES

Figure 2.1: Robot arm and control console.....	11
Figure 2.2: Schematic of control loop.....	13
Figure 2.3: Forward and inverse mappings.	15
Figure 2.4: Controlling the plant and emulator in parallel.....	16
Figure 2.5: Enhanced control console.	17
Figure 2.6: Bioreactor.....	24
Figure 2.7: Neural controller for the bioreactor.....	25
Figure 2.8: Training the emulator.....	26
Figure 2.9: Training the neural controller with an emulator.	27
Figure 3.1: Pseudo-closed-loop control.....	31
Figure 3.2: Schematic of musculoskeletal emulator.	33
Figure 3.3: Left and right eye views of a tetrahedron.	40
Figure 3.4: 'Contextron' processing unit.....	41
Figure 3.5: Effects of movement on retinal projection.	42
Figure 3.6 Real and imagined rotation.	43
Figure 3.7: Sensorimotor feedback apparatus.....	44
Figure 4.1: Plot of some dynamic variables during a typical arm motion.....	59
Figure 4.2: Schematic mapping between physical parameters and analog model..	60
Figure 4.3: Blocks from the block balancing task.	62
Figure 5.1: Schematic of brain emulating the external world.....	81
Figure 5.2: Schematic of brain emulating an agent as emulating the world.....	82
Figure 6.1: 'Jerry understands himself.'.....	94
Figure 6.2: 'Jerry thinks everyone understands himself.'.....	95
Figure 6.3: 'The woman with the hat ate salmon.'	99

Figure 6.4: 'What did the woman with eat salmon?'	100
Figure 6.5: 'What did the woman with the hat eat?'	101
Figure 6.6: 'I drink two rum and cokes with a twist of lemon every day.'	102
Figure 6.7: 'I drink every day two rum and cokes with a twist of lemon.'	103
Figure 6.8: Schematic of 'on.'	109
Figure 6.9: 'Cat on mat.'	109
Figure 6.10: 'Go,' 'gone' and 'away.'	110
Figure 6.11: Hierarchical scene segmentation.	123
Figure 7.1: Cummins' Tower Bridge.	151
Figure 7.2: Generalized tower bridge.	156
Figure 7.3: Tower bridge with non-real target domain.	158
Figure 7.4: Two interpretive projects.	160
Figure 7.5: Interpretation of Twin Earth example.	162
Figure 7.6: Imperfect interpretation and opacity.	170

ACKNOWLEDGMENTS

In addition to the great intellectual debts I owe to many researchers in a number of fields, there are a number of people who have aided me, in one way or another, in the completion of this project. Robert Hecht-Nielsen first presented the ideas to me that began this project. I learned more than just neurocomputing from his course and from the number of conversations we have had. As CEO of HNC, in addition to teaching graduate courses in electrical and computer engineering at UCSD, he has huge demands on his time, and yet he has graciously made time to have a number of long conversations with me about a variety of issues relating to this project, on topics from theories of brain function to issues in engineering generally.

Adrian Cussins has had much more influence on my thought than he suspects, and more than is evident in this project. Earlier drafts of this dissertation included a chapter on metaphysics that manifested that influence, but for purposes of continuity I elected to omit that chapter from the final project. He has made me a better philosopher than I would have been.

Vilayanur Ramachandran was kind enough to allow me to participate in his neuropsychology laboratory for three years even though the amount of time I was able to devote to the lab in hands-on projects was much less than is usual for a lab participant. I had the opportunity to learn from a brilliant experimentalist that amazing revelations are sometimes sitting right under your nose, waiting patiently for you to look at them. His insight, knowledge, and knack for transcending the conventional has been a great inspiration to me.

Ron Langacker has painstakingly developed what I consider to be the most plausible, powerful and elegant theory of human linguistic competence extant. Not only has his work convinced me that language is intimately linked to, and provides a window on,

many other facets of cognition, but he has patiently read and provided crucially helpful feedback on drafts of Chapter Six.

A number of others have helped me in one way or another: by providing valuable feedback on earlier drafts, by forcing me to be less unclear, by offering encouragement, ideas, or inspiration, or by providing opportunities that I would not otherwise have had. These include, but are not limited to Andy Clark, Gary Cottrell, Gilles Fauconnier, Bob Gordon, Peter and Vicki Grush, Valerie Hardcastle, Susan Hibbs, Bill Hirstein, Jordan Hughes, George Lakoff, Megan Lauppe, Phillip Kitcher, Nili Mandelblit, Mary Powers, Joe Ramsey, Georg Schwartz, Oron Shagrir, Micheal Wedin, and Dorene and Dan Wetzel.

Finally, but certainly not least, my debts to Paul and Pat Churchland are countless. Paul's *Scientific Realism and the Plasticity of Mind*, and Pat's *Neurophilosophy*, both of which I read as an undergraduate, were my guiding philosophical inspirations. So much so that the only graduate program I applied to was UCSD's. The lofty standards they set as intellectual inspirations when I was an undergraduate, they matched as mentors during my graduate career. They have provided an environment in which interdisciplinary work is encouraged and where I could follow my interests. They helped in a number of ways (often going far beyond the call of duty) with funding and other administrative issues. When I needed space and freedom, it was there, and when I needed feedback and encouragement, it was there as well. If philosophy really is an honest search for enlightenment, a free investigation into basic questions of existence and human nature, if it is a love of intellectual curiosity unshackled by methodological pretenses and posturings, then no one deserves the title of Philosopher more than Paul and Pat Churchland.

VITA

February 2, 1965 Born, Klammath Falls, Oregon

1990 B.A. (Philosophy) University of California, Davis

1995 Ph.D. (Cognitive Science and Philosophy)
University of California, San Diego

PUBLICATIONS

Grush, R. (1994a) 'Motor models as steps to higher cognition' Behavioral and Brain Sciences 17:2:209-210, commentary on Jeannerod, M. (1994) 'The representing brain - Neural correlates of motor intention and imagery' Behavioral and Brain Sciences 17:2:187-202

Grush, R. (1994b) 'Beyond connectionist vs. classical AI: A control theoretic perspective on development and cognitive science' Behavioral and Brain Sciences 17:4:720 commentary on Karmiloff-Smith, A. (1994) Precis of Beyond Modularity: A developmental perspective on cognitive science Behavioral and Brain Sciences 17(4):693-706

Grush, R. and Churchland, P.S. (1995) 'Gaps in Penrose's Toilings' Journal of Consciousness Studies

FIELDS OF STUDY

Major Field: Philosophy

Studies in Philosophy of Mind
Professors Paul Churchland and Patricia Churchland

Studies in Philosophy of Language
Professor Adrian Cussins

Studies in Cognitive Linguistics
Professors Ronald Langacker and Gilles Fauconnier

Studies in Neurocomputing
Adjunct Professor Robert Hecht-Nielsen

Studies in Neuropsychology
Professor Vilayanur Ramachandran

ABSTRACT OF THE DISSERTATION

Emulation and Cognition

by

Rick Grush

Doctor of Philosophy in Cognitive Science and Philosophy

University of California, San Diego, 1995

Professor Paul M. Churchland, Chair

I explain a strategy, called model-based control, which has proven useful in control theory, and argue that many aspects of brain function can be understood as applications of this strategy. I first demonstrate that in the domain of motor control, there is good evidence that the brain constructs models, or emulators, of musculoskeletal dynamics. I then argue that imagery, motor, visual and otherwise, can be supported by these emulatory mechanisms. I argue that the same apparatus to understanding aspects of psychological development, including the development of theory of mind. I then show how features of linguistic competence can be addressed with the same mechanisms. Finally, I develop a semantic theory applicable to these emulators.

Chapter One: Preliminaries

*Sometimes one must risk error
in order to find truth.*

William James

1.1 Introduction

I think of philosophy as the attempt to grapple with a family of deep and significant questions. When I call these questions deep and significant, I mean that their putative answers promise either justify or to change our picture of the universe and our picture of ourselves as human beings, as thinkers, as moral agents. To this extent I think that most people have probably dealt with philosophical questions at one time or another, and many of the sciences of human behavior address philosophical issues as well. Philosophy as a discipline also approaches such questions, and what makes academic philosophy so powerful, in comparison with the armchair musings of the average person, is that it has a history of well-known questions, well-argued positions, and established methods to draw upon. What makes the philosopher valuable is that he has mastered this history and methodology and is able to bring it to bear on current issues and debates. The terrain of philosophical thought is strewn with pitfalls and obstacles, and the professional philosopher has gone over much of this terrain many times.

But this very advantage can also be philosophy's greatest disadvantage. There is a tendency, especially among professional philosophers, to *identify* philosophy with *this* history and *these* methods for approaching those questions, arguably to the point where the deep and significant questions themselves are upstaged. I say these things because this essay will address questions that I take to be central philosophical questions, but will do so in a way that sometimes departs from established philosophical methodology. For example, nothing prototypically philosophical will appear until Chapter Seven. The first five

substantive chapters will deal instead with issues in control theory, sensory-motor integration, psychology and linguistics. This divergence from philosophical custom is not complete, however. Chapter Seven and parts of Chapter Five are plausibly locatable within current philosophical concerns and practices.

I depart from established philosophical practice (and from standard cognitive science practice, and from psychological practice, etc.) because my purpose is to examine anew the nature of human thought and cognition. I hope to sketch a view of cognition that is applicable to many levels of the cognitive enterprise, from the vestibulo-ocular reflex to choosing a retirement fund, a view which makes sense of certain metaphysical doctrines about mind and reality as well. It will be my contention that if these different levels of cognition and mind are examined in the right way, with the right sort of theoretical machinery, a coherent picture can be seen where before there were only fragments separated by noise. Thus my project is inherently and essentially cross-disciplinary. I do not think that the picture I am peddling could be made to look nearly as attractive if confined to just one area of research, such as developmental psychology, in much the same way that one cannot make a Necker Cube flip (or even be a cube) by staring exclusively at one vertex. To continue the analogy, I hope to take the reader through a number of vertices of cognition, in order to get the terrain to 'flip' into a new, and hopefully more enlightening, configuration.

The danger in any such pursuit is that as scope increases, resolution decreases, and a project as bold (foolhardy?) as this could easily degenerate into vagueness and hand waving. I will attempt to avoid this outcome in two ways. First, at several points I make use of the theoretical machinery I develop to address specific narrow issues. For example, Chapter Five will address issues in the development of the theory of mind, and Chapter Six uses that apparatus in the context of language use to provide insight into wh-extraction and

heavy-NP shift. These diversions serve both to sharpen the theoretical apparatus and to demonstrate that it can cope with concrete issues in a fresh and illuminating way.

The second, and more important, way in which I will avoid the charge of armchair hand waving is by assimilating substantial results from researchers in the various disciplines considered. Each of these disciplines boasts a number of important members who view their work as a reaction against the orthodoxy, and whose work is compatible with the framework I develop. This includes, for example, the treatments of linguistic phenomena found in Langacker and Fauconnier, the developmental psychological work of Karmiloff-Smith and Perner, theoretical neurophysiology of motor control in Kawato and Ito, theories of neurophysiology of perception and imagination as in Llinas and Mel, etc.

To the degree that I am able to assimilate these various results, the original contribution of this work consists in its construction of a unifying meta-theory that promises to provide some rhyme to others' reason.

1.2 Brief history of this project

In the winter quarter of 1992 I was involved in the neuropsychology lab of Vilayanur Ramachandran. At the time the lab was focusing on aspects of neural plasticity, and of the phenomenon of perceptual 'filling-in'. The relationship between filling-in and imagination I found intriguing, but I was not satisfied that any of the explanations or theories with which I was familiar were adequate. At the same time, I was interested in reasoning, and especially long-term planning. As a student of Paul Churchland, one is obligated to take connectionism quite seriously, and yet one area where classical models of cognition seemed to have an intuitive edge was reasoning and planning in the absence of action.

At the same time I was taking Robert Hecht-Nielsen's year-long course in neurocomputing. In the second quarter of this course, we were covering recurrent networks, and the topic turned to using such networks for control purposes. The subject of one lecture was applications to model-based control. This is a technique whereby a model of the controlled system is used in various ways to help the controller control the plant. Hecht-Nielsen's hour and a half lecture, though focused on engineering applications, seemed to me to provide the key to understanding many of the phenomena connected with brain function that I had been wrestling with.

I decided that this would be my dissertation project, to try to show how model-based control could be used by the brain, at a number of different levels, to solve a number of problems. Over the next few weeks, I had worked out rough provisional ideas about how this strategy might be applied to various levels of brain function, specifically in motor control, mental imagery, planning, and language comprehension. The problem with the project, however, was its scope. In order to demonstrate that a certain control strategy is in use at a number of levels of brain organization, one must necessarily learn about, in a fairly detailed way, these levels of brain function and organization. I was confident, however, that I could at least make a plausibility argument in each of the domains under consideration, even if the prospect of detailed and convincing analyses was unrealistic. The promise of a degree of conceptual unification would offset the lack of extended, detailed argument within any given area.

I began my research in motor control, and found quite quickly that a number of researchers within that area were in fact using one or another variant of this control strategy to explain certain phenomena. They were, in effect, doing my work for me, and in a more complete and rigorous manner than I would have been able to myself. I had the luxury, then, of assimilating their work into the overall framework I was trying to defend. The good news, however, did not stop with motor control. In fact, in each area I examined,

there turned out to be a number of researchers whose work was compatible with the framework I was developing. Johnson-Laird's 'mental models' account of reasoning and language comprehension, Langacker's Cognitive Grammar, Mel's simulations of visual imagery, and Fauconnier's 'mental spaces' theory of opacity are several examples.

These discoveries changed the nature of the project slightly. Now freed from the obligation to do a tremendous amount of independent research within each of these fields, I could construct instead a more detailed overall model, and try to show how the work of these researchers, in these different fields, fit together within that model. And that, essentially, is this project. Its scope is considerably larger than is typical for dissertations, but that is necessitated by the thesis, which is that a certain sort of control strategy is applicable to understanding brain function *at a variety of levels of organization and complexity*. And though it is false to say that there is no original contribution in detail to any of the specific domains (there is, for example the treatment of heavy NP-Shift and c-command in Chapter Six, and the solutions to Putnam's and Burge's content attribution problems in Chapter Seven), the bulk of the original contribution lies in providing a framework and common vocabulary within which the work of researchers in different disciplines can be seen to cohere.

1.3 Survey of the Current Project

Having made these brief introductory remarks, I think that the most useful way to flesh out my approach to the questions raised is to simply provide a brief summary of the entire project.

Chapter Two: Emulation and Control

I begin by introducing the notion of emulation, which will serve as the foundation for the entire project. It is a notion I have more or less commandeered from control theory, and is similar to that discipline's notions of a *system identification* or *forward model*. At a first pass, an emulator is an entity that mimics the input/output operation of some distinct target system. For example, a flight simulator is a sort of emulator, as it closely matches the input/output operation of an aircraft, where inputs are command signals from joystick and throttle, etc., and the outputs are instrument readings and visual scene (cockpit window in the case of real airplanes, generated graphics in the case of the simulator).

In this chapter I develop a thought experiment that has a human operator in charge of controlling a large robot arm for the purpose of performing grasping operations. Using this example, I will make the important distinction between inverse and forward mappings, and try to show the many interesting uses that forward models (emulators) can have in such control problems. For example, if one operates the real target system and the emulator in parallel, one can use the emulator's outputs as a check on the real system's sensors and instrument outputs. Furthermore, one can run the emulator by itself (without the target system) to perform 'what if' experiments before executing them with the target system.

Chapter Three: Perception, Imagery, and the Sensorimotor Loop

This chapter will review some of the more substantial evidence that the brain really does employ emulators for a variety of purposes -- specifically, in motor control, mental imagery and perception. The focus is on these 'lower' cognitive functions for two reasons. First, most (but not all) of the hard evidence for the explicit use of emulators comes from these areas, as they are much better understood neurophysiologically than 'higher' cognitive functions. Second, I want to make the case that phylogenetically higher functions

can be seen as adaptations of these 'lower' functions, and thus that the link between Sensorimotor integration and cognition is tighter than might have been supposed.

The first example will be from motor control. Specifically, making use of the work of Ito and Kawato, I will examine certain circuits in the cerebellum whose purpose seems to be the emulation of aspects of musculoskeletal dynamics. The control of very fast voluntary movement faces the difficulty that proprioceptive information from the controlled periphery is transmitted relatively slowly (limited by axon conduction velocities) back to the motor centers, and if motor signals are generated on the basis of old information, oscillations and instabilities can develop. However, an emulator, using efferent copy signals, can provide immediate 'mock' proprioceptive information, information which, if the emulator is a good one, will be the same as the real proprioceptive information that the periphery will eventually generate.

I will then turn to some work in mental imagery which suggests that such imagery is, in effect, simulated perception. I will show that the execution of this simulation requires the use of emulators. I will discuss a computer model of mental imagery done by Mel, which explicitly employs emulators.

Chapter Four: Emulation, Representation and Learning

Emulators, as I develop the notion, are neither connectionist nor 'classical' architecturally, though, of course, an emulational architecture can be *implemented* by either connectionist or classical hardware. In this chapter, I focus on issues in development and learning, and attempt to account for some interesting data in terms of the construction and articulation of emulators. An emulator mimics the input/output function of some target domain, perhaps the external world (we can thus have a 'reality emulator' in our heads, a sort of world model). But there are multiple ways to emulate something. Most basically, one can construct a lookup table of past input/output instances. This might be accurate, but

it will generalize poorly, if at all, to new cases. What one would like is an analog model, where different aspects of the target domain are 'separated out' and can be treated more or less independently from other aspects of the model (articulants, as I shall call them).

This, I think, has interesting connections to the work of Annette Karmiloff-Smith, who offers theories of psychological development which she takes to be neither connectionist nor purely classical. I hope to show how her results fit naturally into the emulational framework (hereafter ETM, for Emulational Theory of Mind). The key phenomenon here will be that of Representational Redescription, a process whereby a representation or capacity gets further articulated in ways which make it more generally applicable, for example to novel situations.

Chapter Five: The Mind/Body Solution

A good model of the behavior of physical objects will make appeal to inertial forces, frictional forces, masses and a host of other ideas -- perhaps in a not very well-articulated manner, like Aristotelian physics, but possibly fairly successful within a certain range. A giant lacuna in such a model will be the behavior of animals and persons, who manifest self-initiated movements, and whose behaviors are straight-forwardly predictable via intuitive physics only occasionally. A solution to this problem is to treat certain physical objects as subject to representational/psychological description as well as physical description.

Drawing on the work of Josef Perner I will argue that this ability results from a recursive application of emulation -- that is, certain entities in the internal emulation are themselves represented as capable of internal emulation of some sort, they are represented as representers.

Chapter Six: The Grammar of Thought

If the brain does in fact operate by heavy use of emulators, many of which model the dynamics of the external world via independently addressable articulants, we might be able to account for the semantics of natural language in procedural terms. That is, expressions of natural language can be viewed as instructions, or procedures, for constructing an internal model or emulator.

This chapter will apply ETM to some linguistic phenomena, specifically wh-extraction and heavy NP-Shift. The idea here is that if natural language works by constructing emulators, then cognitive constraints on what can and cannot be maintained as an articulated emulator ought to be reflected linguistically, as, for example the ungrammaticality of certain sorts of sentence. And alternately, information about what sorts of sentence can and cannot be processed ought to provide clues as to the ability of the brain to maintain certain sorts of emulators.

My work in this chapter will be greatly expedited by the use of Ronald Langacker's Cognitive Linguistics framework, which provides a useful vocabulary for dealing with linguistic phenomena (and which I take to be entirely compatible with ETM), as well as a theory of attentionally mediated segmentation developed by von der Malsburg.

Chapter Seven: Semantics

This chapter begins the serious dive into substantive philosophical issues. I have, in previous chapters, constructed a theory of cognition that places emulators at its core. It was taken for granted that the target system that the higher-level emulators emulate is the real world, or, more accurately, aspects of the real world. The question then arises, are emulators best viewed as neurally implemented *models*, whose articulants get their meaning because they stand for some entity in the target domain (the world)? Or is the semantics of

emulators best viewed as a sort of conceptual-role semantics, the meaning of the separate entities and articulators of the emulators being invested with meaning as a function of their dynamic interaction with other such entities? The point of this chapter will be that the story must be much more complex than either of these two alternatives -- each is, in a sense, both right and wrong.

The investigation begins by exploring some of the more obvious theories of content, causal theories and conceptual role theories. The point will be to get a feel for some of the more important requirements that a theory of content must satisfy, and to get a feel for some of the ways in which different theorists have tried to satisfy these constraints. With this background established, I go on to introduce Rob Cummins' Interpretational Semantics, which, though not correct, provides a convenient starting point for constructing a more adequate semantic theory that makes central appeal to emulation and interpretation.

Chapter Two: Emulation and Control

2.1 A control thought-experiment

Let us begin by examining the following control problem: An operator in a factory has the job of using a large robot arm to perform certain grasping and moving operations. The arm has two joints, a shoulder and an elbow, each with one degree of freedom, such that both degrees of freedom, and thus the arm's range of motion, lie on a plane (see Figure 2.1). For each joint, there are two hydraulic effectors that torque the joint in different directions, much like agonist and antagonist muscle pairs in real musculoskeletal systems. The only direct control the operator has over the arm is through these four effectors. Furthermore, the arm is equipped with sensory equipment to provide the operator with information about the state of the arm. First, there is a video camera mounted above the arm, which feeds to the control room. Second, there are sensors on each of the joints which measure the angle of the joint (these are rough analogues of stretch receptors).

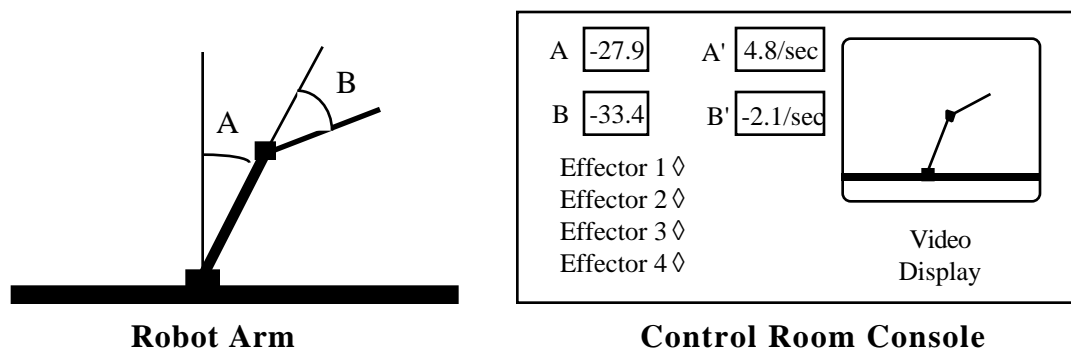


Figure 2.1: Robot arm and control console.

The operator, we will assume, does not have direct visual contact with the arm, but instead operates it from a control room.¹ This control room has a console with the following features. First, there are four toggle switches, one for each effector, which turn the effector on and off. There is a video display, which shows the image from the camera mounted above the arm. Finally, there are sensor readouts which display the angle of each joint, and perhaps its angular velocity as well (the latter can be calculated from the former assuming a continuous reading, or sufficiently small sample interval). We will assume that the operator's task is, when given a target or goal location, to manipulate the arm so that the hand comes to rest at that goal location. The target location might be specified in any number of ways: joint angles, Cartesian coordinates on the plane of motion, or whatever.

A moment's reflection shows that the control problem faced by the operator is not trivial. For example, the same shoulder effector command will have different net effects on the shoulder angle depending on the angle of the elbow, because the angular inertia of the arm as a whole depends, in part, on the relative position of the forearm. Another complication is that, due to 'slingshot effects,' a given elbow effector command will have different effects depending on the motion of the shoulder joint, and motion of the shoulder joint is in turn a function of prior commands sent to both joints. Thus, what effect a given motor command will have depends on previous motor commands in very complex and non-linear ways.

¹ The control room might be thousands or millions of miles from the arm itself. In such extreme cases, which might obtain when operating space exploration equipment from earth, for example, there will be non-trivial delays in the afferent and efferent signals, and these delays can induce problems, including instabilities. These delays are interesting, as we will see later, because the relatively slow speed of axon signal conduction raises similar problems for the central nervous system.

2.2 Inverse vs. Forward Mappings

The first thing we need to do is to characterize the controlled system, or as I shall call it in general, the target system.² In the present example, it will be most natural to take the target system to be the arm and the sensors (where 'sensors' includes the video camera) -- in other words, the target system includes *everything in the control loop between the effector commands leaving the control room and the sensor signals entering the control room* (see Figure 2.2). Defining the target system as some section of the control loop has the advantage that we can treat the target system, for certain purposes, as a black box whose inputs are everything feeding into that section of the control loop, and whose

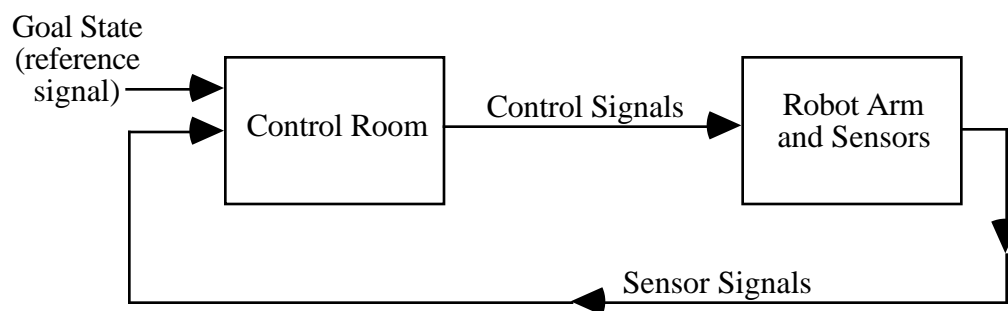


Figure 2.2: Schematic of control loop.

outputs are everything leaving that section of the control loop. Thus target systems are not a natural kind, but are relative to the purpose at hand. For example, from the point of view of the arm and sensors, the control room is an input/output device, and a potential target system -- it takes sensor readings (and perhaps a goal end state for the arm) as input and provides control signals as output (the point is that one cannot just look at a situation and determine what is the target system). So from the point of view of the control room, the

² In the control literature this is also called the plant. Calling the target system the controlled system is somewhat misleading, because there will be cases where the target system is not 'controlled', but merely observed.

robot arm and sensors perform an input/output function: the control room inputs command signals, and the target system outputs sensor readings.

Now that we have characterized the target system in the present example, I want to turn to the control room itself and examine more closely what it does. The control room (which includes the operator) can be thought of as taking as 'input' both a goal end state for the arm, and information (from the sensors) about the initial state of the arm. These are the two pieces of information that we may assume that the control room has no control over at the beginning of each trial. Given this initial input, the control room produces as output a sequence of control signals, and if the controller is successful, the control signals produced are such that they cause the robot hand to end up at the goal end state.

Looked at in this way, the control room performs what is called an *inverse* mapping (the reason it is called an inverse mapping will be clear momentarily). Given an initial arm configuration, if one inputs a goal, the control room outputs an appropriate sequence of commands.³ The arm itself performs what can be called the *forward* mapping. Given an initial state, if one inputs to the arm an appropriate sequence of command signals, it 'outputs' (or ends up in) the goal end state. We can now see why the function performed by the control room is called the inverse mapping: it does the inverse of what the target system does. That is, that target system takes command signals as input, and produces (eventually, if the command signals are appropriate) the correct end state. The control room does the inverse: it takes a goal end state as input and produces a sequence of command signals as output.

³ Notice I said 'an' appropriate sequence, and not 'the' appropriate sequence. As characterized, this inverse mapping is not well-defined, in the sense that infinitely many command sequences will get the arm to the goal (consider: when reaching for a glass, I can do it in any of an infinite number of ways, e.g., bring my hand to my nose and then to the glass, move my hand almost to the glass, bring it over my head, and then back to the glass, etc.). In motor control research, the attempt is made to make the inverse mapping well defined by adding additional constraints, such as time limitations, minimum jerk or minimum torque change constraints, etc.

Notice that if one 'links' the inverse and forward mappings in series, one gets, as one might expect, an identity mapping (see Figure 2.3). Input a goal end state. The inverse mapping (control room) produces a sequence of motor commands. These output commands are fed as input into the forward mapping (the target system), and the target system produces the goal end state as output. So, if the controller is operating properly, it effectively makes the entire system (controller plus arm) act as an identity mapping -- whatever goal state gets input is eventually produced as output by the arm. And it does this by performing the inverse of the target system.

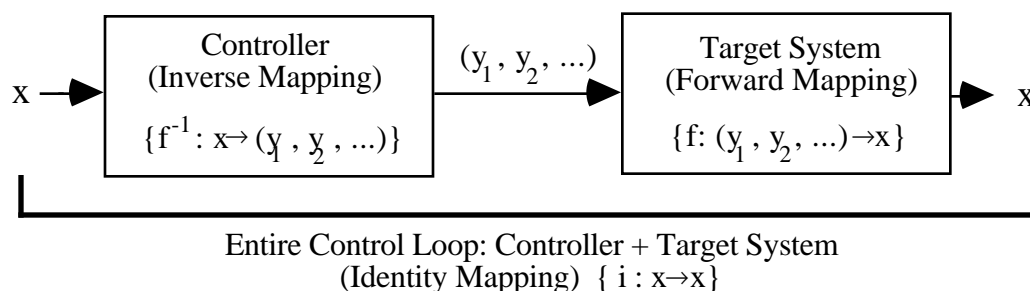


Figure 2.3: Forward and inverse mappings.

2.3 What Is an Emulator?

Finally we are in a position to characterize the centerpiece of this project, the emulator.⁴ In brief, an emulator is a device that mimics the input/output operation of some target system. Since the target system, by definition, implements the forward mapping, the emulator will model the same mapping. Hence, in control theory, emulators are often called 'forward models.'

⁴ Emulators are known by a number of names, including 'model', 'reference model', 'forward model', and even 'system identification'. The term 'emulator' is used in Nguyen and Widrow (1989).

But what does it mean to say that a device mimics the input/output operation of some other device? Well, consider the case of the robot arm. The target system receives, as inputs, command signals in the form of electrical impulses in a wire, say. And it produces sensor/video signals as outputs, again in the form of complex electrical signals. An emulator of this target system would be any device which, for example, could accept as input an exact copy (an efferent copy) of the command signals, and produce an exact copy of the sensor/video (afferent) signals that the real target system would produce as a result of those same commands (see Figure 2.4). A very complex computer program might be able to pull this off.

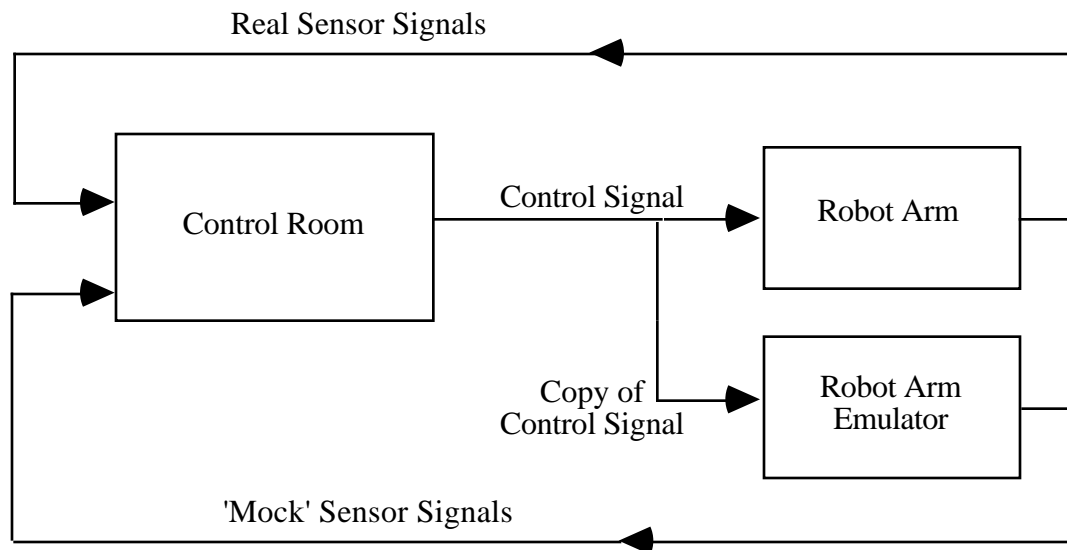


Figure 2.4: Controlling the plant and emulator in parallel.

So let's assume, now, that the control room has some sort of robot arm emulator available, and that its output states are displayed on the control console as well (see Figure 2.5). This setup is exactly like before, except that now there are two sets of sensor/video displays, one for real information about the real robot arm, and a second displaying the emulator's outputs. If the emulator works well, and it is run in parallel with the real system

(i.e. run with copies of the effector commands sent to the real arm) then the information on the two displays should be the same at all times.

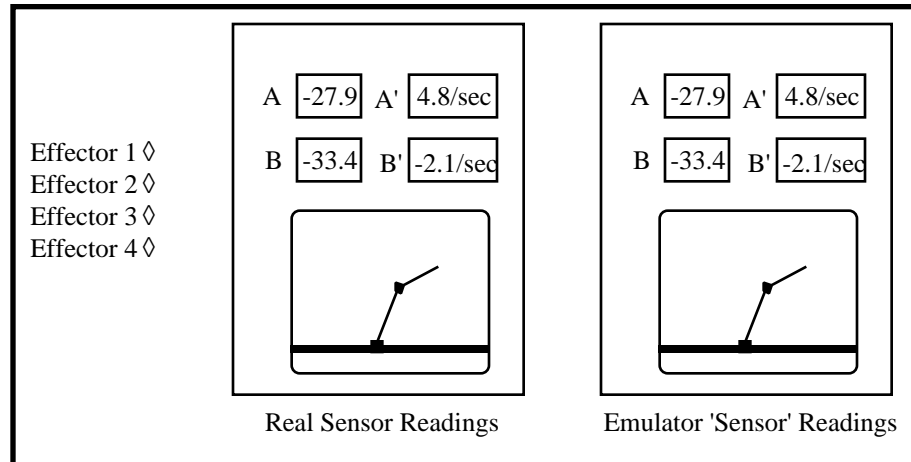


Figure 2.5: Enhanced control console.

Before I continue with the next section, which will briefly discuss some advantages of using emulators, I will run through some other examples of emulation, with the aim of providing a more robust sense of what I mean by the term. A flight simulator, of the sort used to train pilots, is an emulator. In such a case, the target system is an aircraft, and perhaps aspects of the external environment (weather, terrain, etc.). In a real aircraft, a pilot issues command signals via the joystick, throttle, and other controls, and gets 'sensory' feedback in the form of instrument readings and a visual scene (via the windshield). A flight simulator is meant to perform a similar mapping -- it takes command signals and computes output signals, including a mock computer-generated visual display. Notice that in this case, the target system is actually a good chunk of the earth, and the simulator models relevant aspects of that target system, such as what happens when the virtual aircraft runs into a virtual mountainside.

Chess-playing computer programs attempt to emulate a good chess player. Given a board configuration as input (and for more sophisticated programs, perhaps the history of moves in the game as well), the program produces as output a move, much like a real chess-playing human would.

These last two examples bring up some important points. First, concerning the flight simulator, notice that the only aspects of the external environment emulated are those that might make a difference to the simulation. Thus, the fact that mountainsides can be crashed against is emulated, but the fact that weather erodes them is not. The point is that the emulator represents objects under certain aspects, and ignores those aspects that are insignificant for purposes of the emulation. This follows from the fact that in specifying a real system as an input/output system, one necessarily focuses on certain features of that system to the exclusion of others, and this selectivity carries over to the emulation. Emulation-supported representations are context dependent, the context being provided by the domain of emulation.

The chess-playing computer emphasizes another point, which is that some target domains, and hence their emulators, need not be deterministic or even strictly well-defined as input/output devices. Just as I don't think there can be a hard line separating chess players from people who randomly select legal moves, so too there may not be a hard line delineating chess player emulators from random legal move generators. Ultimately what counts as an emulator will be interest relative. So be it.

One final point. I am often asked, after explaining what emulators are, and providing examples of emulation: What *isn't* an emulator? There are two ways to take this question. First, it isn't a requirement for something to be an emulator that it perfectly match the i/o function of the target. Given that, can't anything be characterized as the emulator of something else? Won't the flight simulator be an emulator of my musculoskeletal dynamics, albeit a poor one? Worse, given the latitude possible in specifying just what a

target system is, and what counts as inputs and outputs, we might be able to make the flight simulator a *good* emulator of my musculoskeletal mechanics, so long as we characterize the inputs and outputs in the right (i.e., wrong) way. To this objection I must simply raise again a point made earlier, and then punt. Emulators are interest relative, they are not a natural kind and I don't think necessary and sufficient conditions for emulation can be provided. Ultimately I'm not trying to provide definitions, reductions or foundations, but rather a way of looking at brain function, and to that degree I'm content with a rough, intuitive, 'I know one when I see one' characterization of emulators.

Another way of taking the question is, on theoretical grounds, what is the contrast class for emulators? The answer to this is that the contrast class includes, at least, inverse models, or inverse mappings, and of course the real target system(s). Interestingly, the contrast is very similar to the contrast between prediction and control. An emulator allows one to predict the behavior of a target system, but does not, by itself, allow one to control the target system.

2.4 Benefits of Emulation

Now that we have both developed an example of a control problem and given some idea as to exactly what an emulator is, and enhanced the control room (as in Figure 2.5) with the *prima facie* redundant and unnecessary emulator readout, we turn to the usefulness of emulation. This section will cover several such uses. It is my hope that the reader, while going through these examples, keeps in mind some problems faced by brains in controlling bodies, and can see some of the solutions that emulators offer to control problems as shedding some light on how brains might deal with similar difficulties.

Sensor check, enhancement and replacement

Suppose that the emulator is being run in parallel with the real robot arm, and is accordingly providing duplicate sensor/video information. This duplicate information can be used as a check on the robot arm sensory apparatus, in that any sudden significant divergence between what the real sensors say is happening and what the emulator says should be happening might indicate that there is a sensor failure, or that something is interfering with the normal operation of the arm (a hydraulic effector has failed, for example). Without some idea of what normal sensory feedback *should* be, such malfunctions might go unnoticed, and serious errors might result. This point deserves emphasis. We are so good at predicting consequences of actions, that this very act of predicting often goes unnoticed. Unless there is *some* expectation, malfunction cannot be detected. Emulators can provide a source of expectation. (Of course, rapid divergence might also indicate that the emulator is imperfect. Even so, the knowledge that something may be awry with the target system can alert the operator, and counsel the appropriate cautions.)

Similarly, suppose that the real sensors operate only intermittently, or that the video camera is time-shared between several different systems. No problem: just use the emulator's readouts when necessary. In fact, once the emulator is up and running, one can do without some, or perhaps all of the sensors (if the emulator is very good), and just control the system by way of emulator-produced 'imagery.'

Feedback control under transmission delays

Another benefit of emulation is seen when the controller and controlled system suffer transmission delays in communication. Such transmission delays might occur when the controller and the system are very far apart (as is the case with remote control of space probes), or when the transmission lines are just slow (as is the case with axons). To see

what the problem is, consider the difficulties that might emerge with signal delays. The operator of the robot arm, let us suppose, watches the hand on the video monitor as it approaches the goal position. Just before she sees it reach the goal she sends a control signal to stop the hand. In reality, though, the hand had already passed the goal state when the operator issued the command, because the sensory information she was using was a few seconds old. But there is also an efferent transmission delay, and so the arm has traveled even further past the goal by the time the effectors are engaged. The operator sees the hand continue past the goal, and continues with the effector command to reverse direction. Finally she sees the hand stop and move back towards the goal, but of course by the time she sees this on the screen, the hand has accelerated greatly, and could very well be back on the other side of the goal, speeding away from it rapidly. Depending on various details, the system can oscillate or become unstable.

One possible solution to this problem is for the operator to use the feedback from the emulator to guide her effector commands. Assuming that the emulator is in the control room itself, and that it operates quickly, the operator can have immediate feedback from her control signals, and use that feedback to guide the arm appropriately. This strategy can effectively eliminate transmission delay problems (though of course it introduces other potential problems).

Control room imagery for planning and 'thought experiments'

Suppose that the operator is asked to make a certain maneuver with the arm, one that is dangerous, for example, in that it brings parts of the arm near fragile obstacles. The operator may wish, before attempting the move, to first disconnect the real arm control, and try the maneuver with the emulator to see if it will work. If it does, then the operator might have more confidence in the real arm's capacity to execute the same moves safely. Relatedly, one can perform thought experiments with the emulator, such as: Can we

perform all the functions we have to while limiting the elbow to half the range of motion?⁵ One can use such imagery to try several different solutions to a problem, and see which works best before committing time and resources to (and accepting the liabilities of) a real attempt. In situations where there are many obstacles, and perhaps more degrees of freedom in the manipulator than two, it may not be immediately clear if there is a way to move the arm to the desired position without hitting an obstacle.

Enhanced controller training capabilities

Having a good emulator makes training controllers much easier. The first advantage is that it makes such training safe (this is of course a big benefit of flight simulators - simulated crashes don't kill people or destroy property). Furthermore, such training can be much less expensive (consider the cost of jet fuel). And also training can be much more efficient because with an emulator one can normally dispense with the set-up time usually required with real systems.

Another advantage is that training on an emulator allows one to gain experience with rare or dangerous situations that one would not want to create on real systems for training purposes -- the ability to practice pulling out of stalls without ever actually having to be in a stalled aircraft, for example, or to train a nuclear reactor controller to operate effectively in critical conditions. Finally, if the controller one is training is some sort of automatic adaptive controller, supported by a computer program, then assuming that the emulator is software supported as well, one can run the training much more quickly than on the real system, because there is no need to do it in real time.

⁵ Consider other cases, such as chemical plants, where plant time and chemicals used are very costly. Operators might want to know if they can achieve a certain level of performance with different mixes, or under different conditions, while still maintaining certain safety parameters.

2.5 Emulation and Neural Networks

In this final section I will briefly discuss a neural network implementation of emulation. I include it here for several reasons. First, I think that at some level, perhaps some sort of implementation level, brains operate in a more or less neural-network manner.⁶ Given this, it is important to see how something like a neural network might implement something like an emulator. One way of looking at the story I will be telling is that the Emulational Theory of Mind (ETM) is really neither connectionist nor classical (in fact, it's not really a definite architecture, either), though I think one can see how emulators can be implemented in a neural substrate.

Here we switch examples, from a human operating a robot arm to an artificial adaptive controller operating a bioreactor. The bioreactor presents a non-linear control problem which is often used as a benchmark for assessing control algorithms.⁷ The bioreactor consists of a tank, which contains a certain amount of nutrient-enriched liquid; some living cells, which consume the nutrients and reproduce; a stirring mechanism, which agitates the contents of the tank; and a heating element, which can warm the contents of the tank (see Figure 2.6). The goal is to keep the total biomass in the tank constant (i.e. keep the total number of cells in the tank constant, assuming uniform cell mass). To this end, the controller, human or otherwise, can manipulate three control variables. The first is nutrient inflow, which is simply more cell-free nutrient-enriched liquid. The total volume of tank

⁶ The term 'connectionist architecture' is used to refer to many different things. I think it is unfortunate that it has come to be associated more and more with a handful of specific sorts of models, such as Hopfield networks and basic backprop nets (Cf. McClelland and Rumelhart (1986)). Neural architectures are much more inclusive than these examples suggest. Thus I can agree with Fodor and Pylyshyn (1988) that 'connectionist' architectures are not adequate to account for certain cognitive phenomena, provided by 'connectionist' it is understood that one refers only to the sorts of models and examples they refer to. But there are much more powerful sorts of connectionist architectures, which are not *prima facie* subject to the same objections. Having raised this point, I propose to drop it. I don't, in the final analysis, much care who gets to stick their flag in the phrase 'cognitive architecture.'

⁷Cf. Ungar (1990)

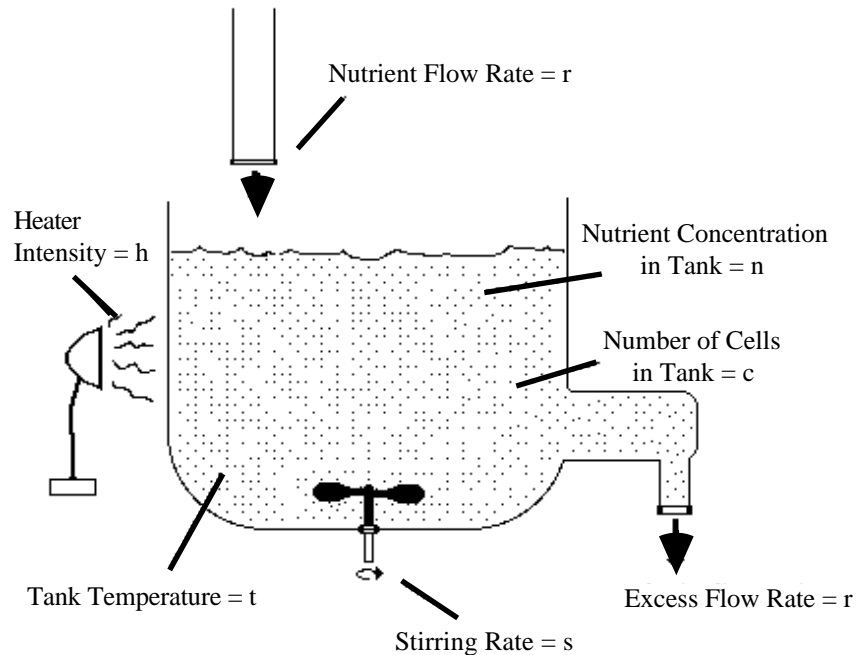


Figure 2.6: Bioreactor.

contents is kept constant by draining the excess at a rate equal to the rate of nutrient-enriched liquid inflow. This draining, of course, results in loss of cells, and thus one could in principle eliminate all cells from the tank through a continuous high nutrient inflow rate, which would effectively flush the tank.

The second control variable is the heater intensity. Cell metabolism, and hence nutrient consumption rates and reproductive rates, will vary with tank temperature. Finally, the stirring rate will affect the ability of cells to absorb nutrients by eliminating areas where cell concentration is high and nutrients have been depleted. Stirring affects temperature uniformity as well.

So now let us suppose that we want to train a neural network to control the bioreactor -- to manipulate the control variables in such a way that the number of cells remains constant (see Figure 2.7). Initially one might assume that this should be easy -- just train a feedforward backprop net whose inputs are the current state variables (number of cells, tank temperature, nutrient concentration) and whose outputs are the appropriate control variables (nutrient inflow rate, stirring rate, and heater intensity). This might sound

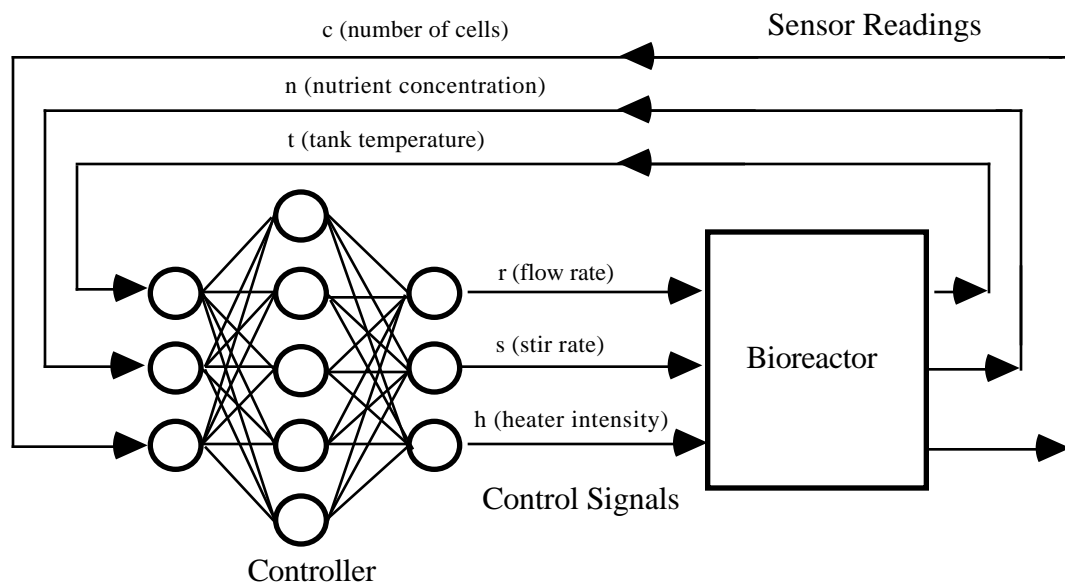


Figure 2.7: Neural controller for the bioreactor.

nice, but it cannot be done as is -- not because the problem is too difficult or because there will not be enough training data. The reason it won't work (unless one uses an emulator, see later) is that in order to train a backprop net, the net must have an appropriate error signal. That is, when the net outputs some control signals, we must know what the correct (or good enough) control signals really are in order to determine the network's error, and then this error must be used to adjust the weights in the net.

But the only error available is the difference between the actual number of cells and the target number of cells. Unfortunately, 'number of cells' is not a network output. The

network outputs are nutrient inflow, stirring rate and heater intensity. Thus any network error will have to be in these terms in order to be useful. But all we know is deviation from target cell levels, and since we cannot backpropagate this error through the bioreactor to get error signals for the controller, it is of no help.

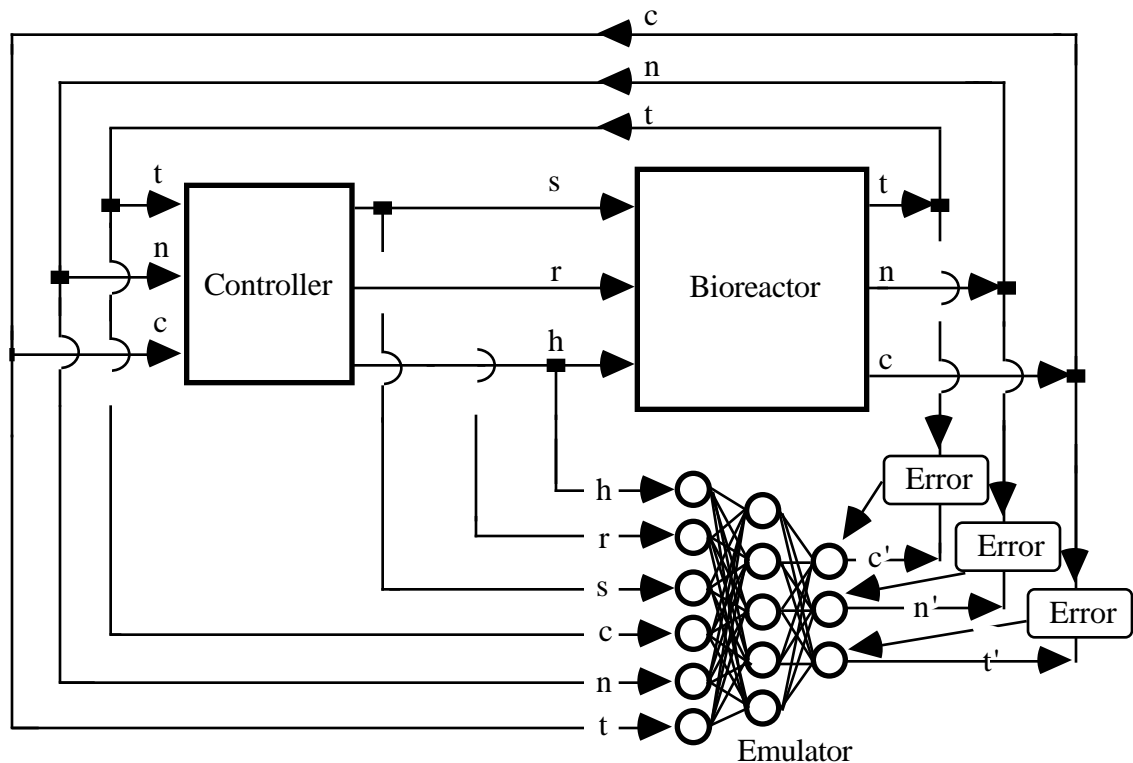


Figure 2.8: Training the emulator.

But what if we could backpropagate this error through the bioreactor to get derivative information that *would* be of use to the controller? Then the problem would be entirely tractable, since any error in target cell levels could be used to determine control signal errors, and these errors in turn could be used to train the controller. Fortunately, this can be done. The key is to first train a neural network to *emulate* the bioreactor (not control it). The emulator can be trained by giving it the same inputs as the real bioreactor, and using any difference between the net's predicted state signals and the real state signals as

the error (see Figure 2.8). The emulator will also need to have inputs specifying the current system state, and these can either be provided through inputs from the real bioreactor (during training) or from recurrent connections from its own outputs.

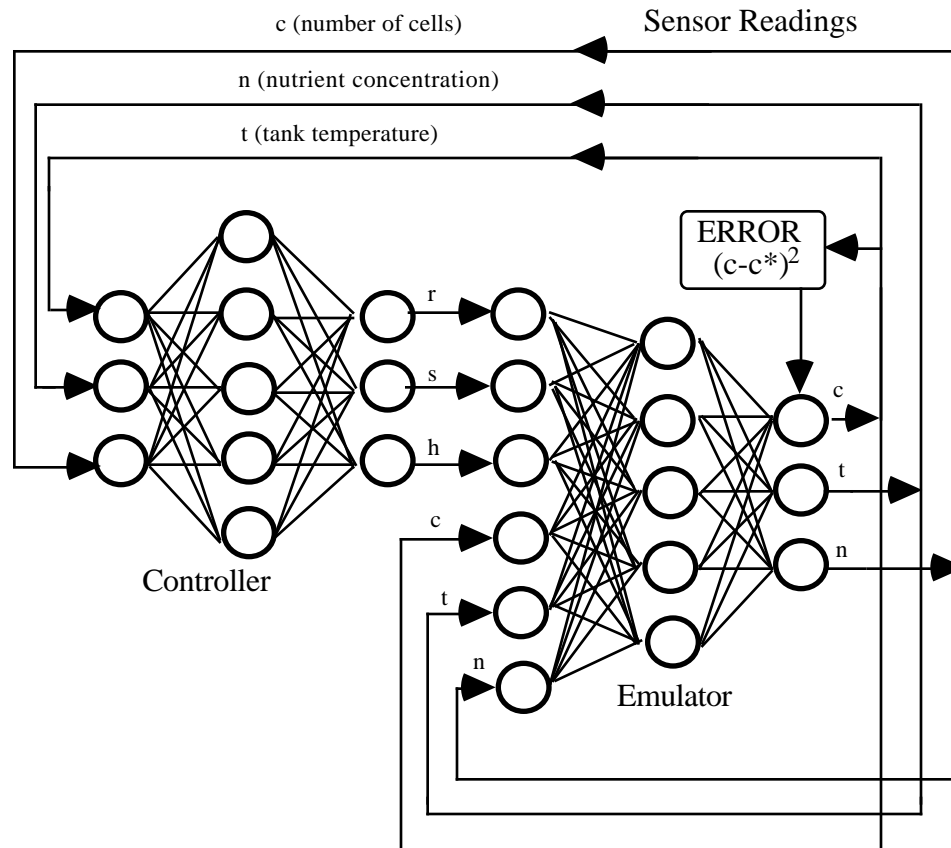


Figure 2.9: Training the neural controller with an emulator.

Thus, once the emulator is trained and operating well, once its input/output operation matches that of the real bioreactor over a wide dynamic range, then its own outputs can be fed into its three state variable inputs, leaving the network with three external inputs (the control variables) and three outputs (the state variables), exactly like the

real bioreactor (thus the emulator and the bioreactor are 'plug compatible'). At that point the controller can be trained to control the emulator, and once it can do this, since the emulator works just like the real system, the controller can be used to control the bioreactor. The advantage in doing this is that the neural network emulator is differentiable -- one can backpropagate through it, and so now the error in the target number of cells can be used to generate an error signal for the controller at each time step. Of course, at this stage, even though one backpropagates the cell-count error through the emulator to get a control signal error, one does not change the emulator's weights, but only the weights of the controller (see Figure 2.9).

This concludes the business of the present chapter.⁸ I hope that, even if the reader doesn't believe everything I say about emulators, at least what I mean by the term is clear enough. The next chapter will examine some issues in human sensory-motor integration and imagery in the light provided by the previous discussion.

⁸ I would like to thank Robert Hecht-Nielsen for valuable discussions of, and lectures on, these topics. I would guide the reader who is interested in these issues to Miller, Sutton and Werbos (1990). The chapters therein, and their bibliographies, are quite rich, and their influence on many of the ideas presented in this chapter is substantial.

Chapter Three: Perception, Imagery and the Sensorimotor Loop

Perception is basically a controlled hallucination process.

Ramesh Jain

Let us begin with the motor control centers of the central nervous system, and make some arguments, in the spirit of Kawato and Ito, that certain circuits in those centers act as emulators of musculoskeletal dynamics. I will then present some arguments to the effect that such emulators drive motor imagery.⁹ Then, I will make the case that all imagery is best understood as emulational in nature. Finally, using some arguments compatible with and supported by considerations put forward by Llinas, I will argue that perception itself is best viewed as a sort of imagery, and hence as dependent, ultimately, on emulation.

3.1 Motor Control

This section will focus on some problems associated with motor limb control. The problem, hinted at in the previous chapter, is that a broad class of motor functions requires feedback faster than it is available from the periphery, and this feedback delay can cause oscillations or instabilities. These movements are fast voluntary movements, as opposed to slow voluntary movements, or cyclical movements like walking or running which might be largely controlled by central pattern generators in the spinal cord. One way that the nervous

⁹ I will use the term 'imagery' or 'mental imagery' as a blanket term to cover all sorts of imagery, including visual, auditory, motor, etc. Imagery is often equated with visual imagery, but I want to keep it as a generic term, and will modify it with 'visual' or 'motor', etc., when necessary.

system can solve this problem is to use emulators of the appropriate system to provide much faster feedback, through processing of efferent copy signals, to the control centers.

3.1.1 Fast, Accurate Motor Control: The Problem

The brain faces many obstacles when trying to execute fast, accurate voluntary movements, many of which are similar to the problems faced by the robot arm operator in the previous chapter as a result of feedback delay. When the brain issues a motor command, it takes some time for the signal to traverse the spinal cord. In the most favorable case, there will be one synaptic relay where the efferent axon from primary motor cortex contacts a motor neuron, and then more delay as the motor neuron axon carries the signal to the muscle fiber. Then there is yet another synaptic transmission at the neuromuscular junction, and finally the muscle responds. The muscle stretch receptors and the Golgi tendon organs must relay proprioceptive information¹⁰ from the limb back to the spinal cord, where it continues, again limited by axonal conduction velocities.

The exact temporal length of this feedback loop is not known, but there is good evidence to the effect that 500ms is an approximate lower bound for any proprioceptive information to be effectively available for feedback control.¹¹ That is, movements taking less than 500ms are executed (at least as far as proprioceptive feedback goes) open loop. Visual feedback is available much more quickly, but it is still subject to significant delays. Furthermore, there is good evidence that trajectory corrections are made *in the absence of visual feedback, at latencies far below the proprioceptive feedback loop time* (see below).

¹⁰ The stretch receptors and Golgi tendon organs are sensitive, roughly, to muscle length and tension at the muscle-tendon junction, respectively. And these values are roughly correlated with joint angle and joint torque, respectively. In addition, as a function of how fast the receptors adapt, some are responsive to the values of these parameters, while others, which adapt quickly, are sensitive to changes in these parameters, and thus can give an indication of their time derivatives.

¹¹ See Denier van der Gon (1988). Of course this does not imply that it takes a full 500ms for such information to reach central control areas. It may be there somewhat faster. The experiments cited modify peripheral signals to determine their effects on fast movements, with the finding that, with movements that last less than 500ms, such modification makes no change to the movement. Given this, changes to the motor signal that occur significantly before 500ms cannot be made on the basis of peripheral feedback.

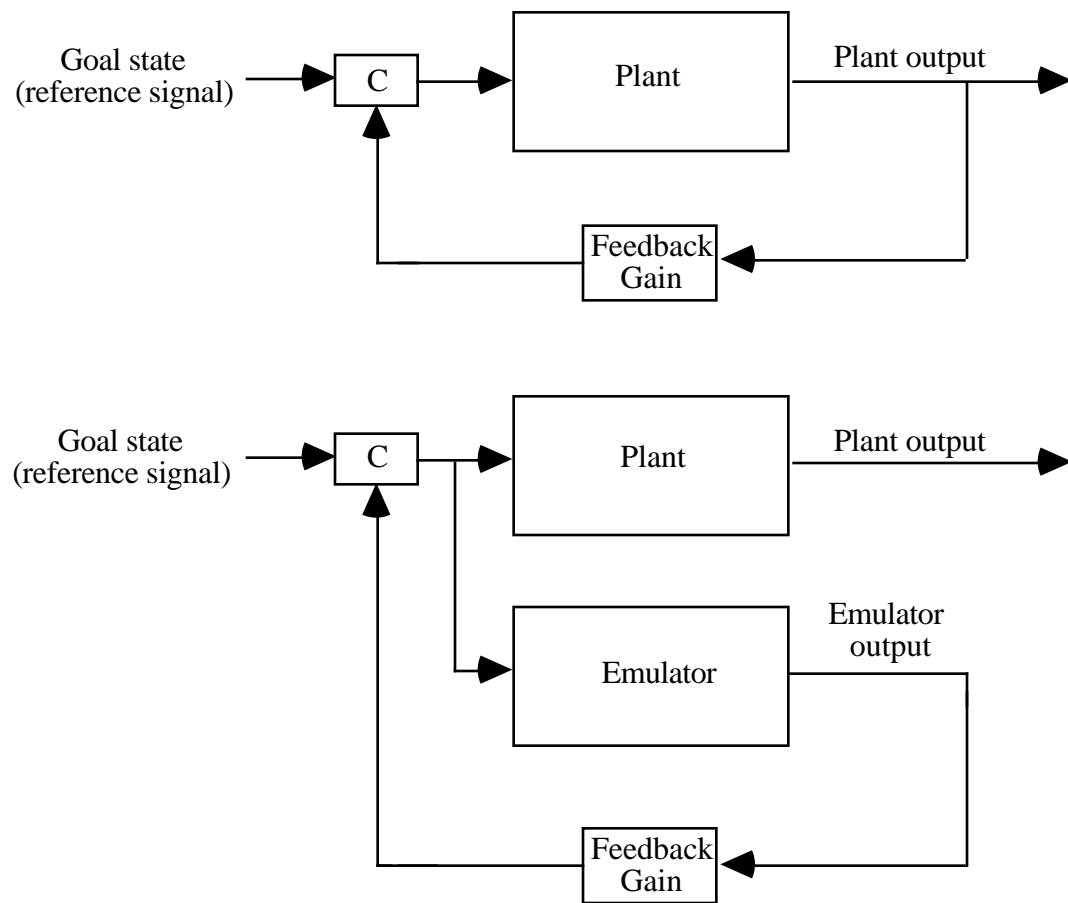


Figure 3.1: Pseudo-closed-loop control.

One way that greater accuracy could be attained in the execution of open loop control is to try to make the task pseudo-closed-loop by using an emulator to process efferent copy information, and using the emulator's output to make adjustments to the motor command. Figure 3.1 shows a normal closed loop feedback control schematic, where output from the plant is used by the controller (C) to adjust control signals. The lower schematic shows a pseudo-closed loop architecture, where efferent copy information is processed by an emulator, and its output is used by the controller to adjust the control signal. The necessity of such emulators of musculoskeletal dynamics for this and related purposes is recognized by a number of researchers in motor control. What is less agreed

upon is the exact location and use of these emulators, though circuits involving the cerebellum and various brain stem and midbrain nuclei such as the red nucleus, pontine nuclei and the reticular nuclei seem to be the odds-on favorites.¹² The remainder of this section will outline the proposal made by Kawato for such an emulator. I choose this example because it is fairly well articulated, and it shows how such an emulator can be of use not only for the control of motor tasks, but also for motor learning (specifically, the acquisition of a good *inverse* model).

3.1.2 Fast, Accurate Motor Control: A Solution

As Kawato et al. write:

The spinocerebellum (vermis and intermediate part of the hemisphere) - magnocellular part of the red nucleus system receives information about the results of the movement ... an afferent input from the proprioceptors, as well as an efference copy of the motor command. *[That is, this system sees a copy of the motor command, as well as the proprioceptive information it leads to. Association of the first with the second amounts to learning the forward model. -RG]* Within this spinocerebellum-magnocellular red nucleus system, an internal neural model of the musculoskeletal system is acquired. Once the internal model is formed by motor learning, it can provide an approximated prediction of the actual movement... A predicted possible movement error ... is transmitted to the motor cortex and to the muscles via the rubrospinal tract. Since the loop time of the cerebro-cerebellar communication loop is 10 - 20ms (Eccles 1979) and is shorter than that of the supraspinal loop, the performance of the feedforward control with the internal model and the internal feedback loop is better than that of long-loop sensory feedback. (Kawato, Fukahara and Suzuki 1987)

Kawato proposes that a specific circuit (see Figure 3.2) performs the function of emulating the musculoskeletal dynamics. According to this model, efferent copies from the

¹² The best and most thorough discussion of these control issues I have seen is Ito (1984). Others who explicitly argue for the necessity of a musculoskeletal emulator are Tsukahara and Kawato (1982), Kawato et al. (1987), Kawato (1989) (1990), Houk (1988) (1990), and Arbib (1981).

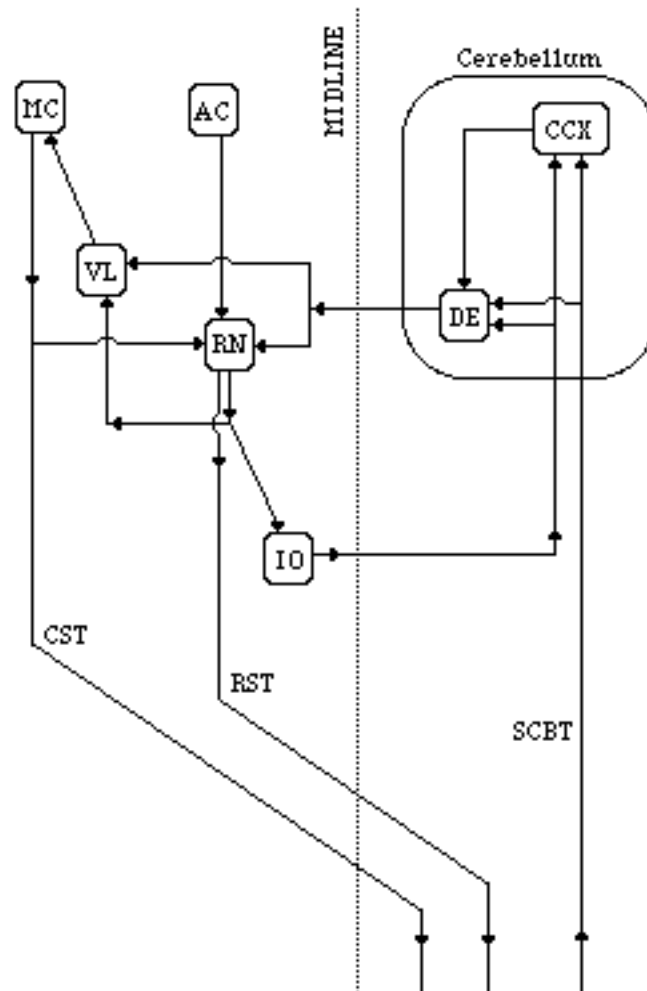


Figure 3.2: Schematic of musculoskeletal emulator. Adapted from Kawato et al. (1987).

cortico-spinal tract are sent to the red nucleus, where they enter a circuit that traverses the inferior olivary nucleus, then to the contralateral dentate and cerebellar cortex via climbing fibers, and returns to the red nucleus via the dentate. This loop embodies the emulation of musculoskeletal dynamics, and provides a predicted position on the basis of the efferent copy. There are two possible ways that this information can lead to an update of the motor signal. First, the desired movement (communicated from the association cortex) can be compared in the red nucleus with the predicted movement, and a correction signal sent via the rubrospinal tract. Second, the predicted movement can be communicated to the motor

cortex via the ventral lateral thalamus, where a correction can be made. Not only does this allow for better feedforward control, but it also aids the learning of motor skills in several ways. First, the emulator circuit provides very fast feedback concerning the effect of a given motor command, and it is plausible to suppose that, at least in some cases, synaptic plasticity is easier to regulate when feedback is immediate. Second, learning of the inverse model can take place in the absence of actual movement. An action can be mentally rehearsed over and over to provide the controller with feedback about the results of its actions.¹³ Notice that these two benefits, greater control and enhanced learning, which are direct results of the use of emulators, have long been attributed to the cerebellum and associated motor nuclei. Neurophysiology texts and articles almost always gloss the function of these structures as 'playing a role in the control of fine movements and motor learning,' but usually do so without any *theoretical* motivation. Emulation gives us exactly this motivation.

At this point it would be nice to see if there are in fact any data to support the claim that Kawato's (or some similar) control structure is in fact used. In a series of rather clever experiments,¹⁴ subjects were asked to make fast arm movements from a start location to a goal location. Subjects were allowed time to see the goal, and were not under pressure to *initiate* the movement quickly (they had time to think about the motion before execution), but were asked to make the movement quickly and smoothly once initiated, and to try to stop the hand as close to the goal as possible. Trials were conducted both with and without visual feedback.

¹³ It has been shown that mental rehearsal of physical tasks facilitates and improves subsequent performance. See Feltz and Landers (1983) for review of extensive literature; also see Yue and Cole (1992), and Jeannerod (1994). See also discussion below in next section. On a more speculative note, it is tempting to wonder if one of the functions of dreaming is to allow inverse models to train on emulators. As mentioned in the previous chapter, a benefit of training on an emulator is that it allows one to train for dangerous situations without assuming real risk, and to train on unusual situations that occur rarely enough to effectively negate any chance of experience with the real thing.

¹⁴ See van der Meulen, Gooskens et al. (1990).

Position, velocity and acceleration analysis showed that the first 70ms of such movements have a great degree of variance, i.e. there is little inter-trial correlation in the distance traveled up to 70ms. However, *after* 70ms, but *before* proprioceptive information is available, and in the absence of visual feedback, corrective adjustments are made such that the total distance travelled in the acceleration phase is closely correlated across trials. (It was shown that visual feedback made a significant contribution only *in the deceleration phase* of the movement.) These initial corrective adjustments during the acceleration phase of the motion, because they were executed in the absence of visual feedback and beneath the long-loop feedback threshold, had to have been made on the basis of a short-loop internal feedback signal, which requires an emulator. Notice, finally, that should anything happen to the emulator/short-loop feedback system, only long-loop feedback would be available for corrective movements, and we could predict that this would result in oscillations or instabilities, especially towards the termination of a movement. As Kawato notes, this in fact is exactly what occurs as a result of cerebellar dysfunction, and is known in the literature as tremor (or specifically, intention tremor).

As a final note, Ghez and Vicaro (1978) and Ghez (1990) have reported neurons in the magnocellular red nucleus of the cat (one of the nuclei implicated by Kawato in the emulator loop) that modulate their activity as a function of the time derivative of force exerted on the limb, but do so *before* the limb is acted on. That is, the activity of these neurons predicts limb jerk. Ghez and Vicaro assume that these neurons are motor neurons innervating the muscle of the limb in question, which is one plausible interpretation. However, another possibility is that these neurons are in fact predicting future limb jerk on the basis of efferent copy signals.¹⁵ Supporting the second interpretation is the fact that the

¹⁵ To be speculative for a moment, it was pointed out earlier that inverse motor control mappings are ill-defined, and that motor control theorists add additional constraints such as minimum jerk or minimum torque change criteria to make inverse dynamics well-defined. Given that dF/dt is directly proportional to jerk, and closely related to torque change, cells predicting dF/dt could play an obvious role in dynamic profile determination, since a profile that minimizes the activity of such units would be a minimum jerk profile.

same authors found that these cells also responded to passive limb movement (recall that an emulator needs not only an efference copy, but information concerning the current dynamic state of the target system). Regardless of which interpretation is correct in this particular case, the ambiguity points out the importance of theory (or lack thereof) in interpreting experimental results.

3.2 Motor Imagery

If the previous section is right, then the brain has circuits that take as input information regarding the current state of the body, as well as a copy of an efferent command signal, and compute as output the state or configuration of the body that will result if those motor commands are successfully executed. In the simplest case this end-state specification will be given in proprioceptive terms. The prediction generated by the musculoskeletal emulator is a prediction of the proprioceptive (as opposed to visual, etc.) state of the limb. This raises an interesting possibility, which is that this very same mechanism can support motor or kinesthetic *imagery*. The only modifications necessary are first, that the motor command itself be inhibited from acting on the musculature; second, that the 'current state' specification of the system be provided via some sort of recurrent pathway; and finally, that the outputs of this emulator be made available to those centers that support normal proprioception.

This section will argue that there is evidence that motor imagery is supported by similar structures that support motor function.¹⁶ The next section will build on this idea, and make the case that all imagery can be similarly explained as a sort of simulated perception. The final section of this chapter will show that if imagery really is simulated

¹⁶ The argument structure here owes much to Jeannerod (1994), who also argues that motor control and motor imagery are supported by similar structures. However, Jeannerod does not recognize that emulators are a necessary part of this puzzle (as I pointed out in Grush (1994a)).

perception, then the distinction between perception and imagination begins to blur in some interesting ways. But first to motor imagery.

One might think of motor imagery as primarily a sensory phenomenon, since its result is imagined sensations. However, if it depends on the processing of motor commands that are inhibited from acting on the musculature, then one would expect motor areas to be active during motor imagery. This is exactly the case. A host of studies¹⁷ have confirmed the following pattern: During many sorts of *overt* motor activity (mostly sorts involving some degree of attention), not only does primary motor cortex show increased metabolic activity, but so do the supplementary motor (SMA) and premotor areas. However, during mental *simulation* of the same movements, primary motor cortex shows no significant increase in activity, *but SMA and premotor areas do show increased activity*. This would suggest that the efferent copy used to compute the predicted movement originates in SMA or premotor cortex,¹⁸ and that during imagery, the normal efferent pathway is inhibited at or before primary motor cortex.

Another line of evidence also suggests that motor behavior and imagery have a similar origin. It has been shown¹⁹ that vegetative functions (such as respiration and heart rate) are affected, at least in part, by central motor commands, as opposed to being driven by, e.g. venous blood CO₂ concentration. For example, heart rate and respiration increase very quickly (within a few seconds) of the onset of strenuous activity, substantially before such activity could result in increased CO₂ concentration. This suggests that such

¹⁷ Some such studies are Decety, Sjöholm et al. (1990); Roland, Larsen et al. (1980); Fox et al. (1987); Ingvar and Philipsson (1977). The study by Decety, Sjöholm et al. (1990) is particularly interesting because not only do they confirm the pattern of rCBF (regional cerebral blood flow) at issue, but they also implicate the cerebellum in imagined motor activity. If the hypothesis of Kawato (and Ito and Houk, etc.) that the cerebellum participates in an emulatory loop is correct, and if I am correct in assuming that such an emulator supports mental imagery, then this increase in cerebellar activity during motor imagination is to be expected.

¹⁸ Jeannerod (1994) reaches the same conclusion, though for different reasons (he is interested not in emulation but in motor imagery). He argues that the signals used for imagery originate in premotor cortex or the basal ganglia.

¹⁹ See Goodwin et al. (1972) and Requin et al. (1991). See also Jeannerod (1994).

vegetative effects are set in motion by central motor commands, or copies thereof. It has also been shown (Decety et al. 1991, Wang and Morgan 1992) that imagined strenuous activity likewise increases heart rate and respiration, and that this increase is proportional to the intensity of the activity being imagined. Again, the suggestion is that overt motor activity and motor imagery are initiated by the same mechanism, but that in imagery overt activity is suppressed, and that this mechanism plays a role in modulating certain vegetative functions.

Additional evidence comes from the sports physiology literature, which has long confirmed²⁰ that imagined rehearsal of some physical activity facilitates its subsequent performance. On the assumption that normal motor performance increases with practice because feedback can be used to adjust the effector sequence, and on the further assumption that motor imagination is just like motor performance except that the emulator provides the feedback instead of the proprioceptors themselves, then the benefits of imagined practice are unmysterious. To the degree that the emulator faithfully reproduces the feedback that the real system would produce, exactly the same benefits accrue.

What I have not provided is any detailed neuroanatomical or neurophysiological model of the mechanics of motor imagery. I have made some suggestions which support the contention that those mechanisms are driven by the same areas that drive motor activity, and that the mock sensations produced are the outputs of an emulator, perhaps the same emulator implicated in motor control in the previous section.²¹ But the bottom line is that the arguments made here are inconclusive, and I will not pretend otherwise. However, I do think that the considerations are plausible, suggestive and interesting. Next, I plan to generalize the story of this section to cover all imagery, especially visual imagery.

²⁰ See footnote 13.

²¹ It is of course possible that there be more than one musculoskeletal emulator, one which is used for control purposes, and one which supports imagery. While the Decety et al. (1990) finding that motor imagery increases cerebellum metabolism suggests that it might be the same, I plan to remain neutral.

3.3 Visual Imagery

The previous section focused on motor imagery, the imagination of the exertion of effort and the resultant proprioceptive sensations such effort would normally produce. But motor commands have effects not only on future proprioception, but on future visual perception as well. When I walk ahead the visual scene changes in interesting and repeatable ways. When I examine a coffee mug by turning it around in my hands, the projection of the mug onto my retina changes, again in at least partially predictable ways. In fact upon reflection there seem to be very few visual scenes that are not changing continually as a function (at least in part) of centrally generated motor commands, whether to the legs, the hands, the neck, or the muscles of the eyes. This suggests that it might be possible to build a visuomotor emulator, one which has as input a current retinotopic projection (and maybe some past retinal states as well) together with a current motor command, and predicts what the next visual scene (or retinal projection) will be.

I will begin this section by describing a fascinating connectionist simulation by Mel (1986), a simulation of a robot that learns to be able to produce visual imagery, including rotation, zoom and pan, by constructing an emulator of its motor-visual loop. Before delving into the details of the model, let's look at why such a mapping really is an emulator. An emulator, recall, is a device that mimics the behavior of some target system, and there can be considerable latitude in what counts as a target. Target systems were defined in the previous chapter as the 'other' side of the control loop. In the case of visual perception, this includes everything between the motor effectors (innervating, e.g., the legs, neck, and eyes) and the visual input (on the retina, or area VI: the exact location

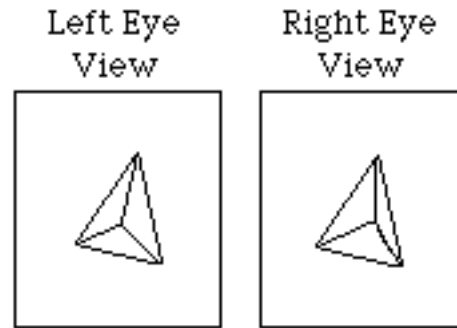


Figure 3.3: Left and right eye views of a tetrahedron. Adapted from Mel (1986).

is not important for the present conceptual point). Thus the target system in this case includes a great deal of the visual world, especially the statistical regularities and other invariances it displays. The important point is that there is at least a partial dependence of the next visual input on the current input and the current motor commands. For example, if I am foveating on a colored square, and I walk forward, the projection of the square on my retina will increase in size, *ceteris paribus*, and this dependence, and many others, can be learned and exploited.

Mel's model does exactly this. The model (which is virtual) has two retinae and several motor effectors, including move forward, move back, move left, move right (I will assume here for simplicity that 'move left' and 'move right' refer to circular motion, such as moving around an object to get different views of it. My exposition of Mel's model will make some other simplifications as well. For greater detail, see Mel (1986)). Each of the two retinae is a 2-D array of processors, and each receives a projection of some 3-D wire-frame 'object,' such as a cube or tetrahedron (these left- and right-eye views are generated by a graphics package: see Figure 3.3). The model, a sort of virtual robot, can also 'issue'²² any of a small number of motor commands, and the retinal projections are recalculated at each time step on that basis.

²² It is not necessary that the robot 'willfully' act in any way. What is required is that the model be able to distinguish the different motor 'contexts', so that it can learn the forward mapping of those contexts. In

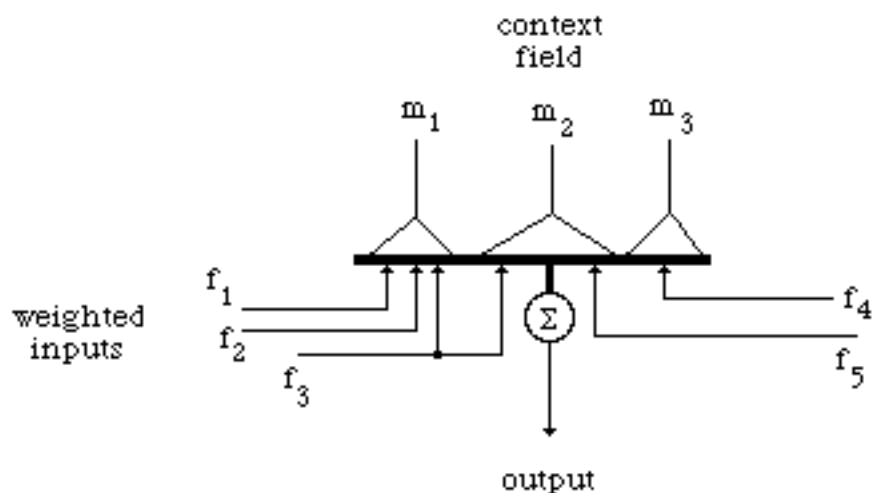


Figure 3.4: 'Contextron' processing unit. Adapted from Mel (1986).

The processing element, which is driven by external input in the normal 'visual' case, receives connections from neighboring processors, (f_1 - f_5). These inputs are gated by 'motion context' inputs, m_1 - m_3 , which gate on those inputs that act as good predictors during the sort of motion (backward, forward, etc.) that they represent.

During the learning phase, the robot simply moves itself around virtual 3-D objects, and learns to predict future 'retinal states' on the basis of the current retinal state and the motor command. The weights connecting the retinal cells to each other and to the motion context cells learn the forward model of the visual-motor transformations of 3-D objects. Each retinal unit projects to all neighboring units within a certain area, and in addition each retino-retinal connection is matched with 'motion context' connections which fire when and only when that motion is being executed (see Figure 3.4). These motion context connections have the effect (after training) of gating the appropriate intra-retinal projections that will act as good predictors in that motion context. For example, during motion directly toward an object which projects onto the retina, the resultant increase in projection size is

animals, presumably, this information is available to the brain because the brain itself issues the motor commands.

manifested as radial motion away from the center of the retina (see Figure 3.5). Thus in the motion context 'move forward,' excitatory connections to units that lie in a straight line

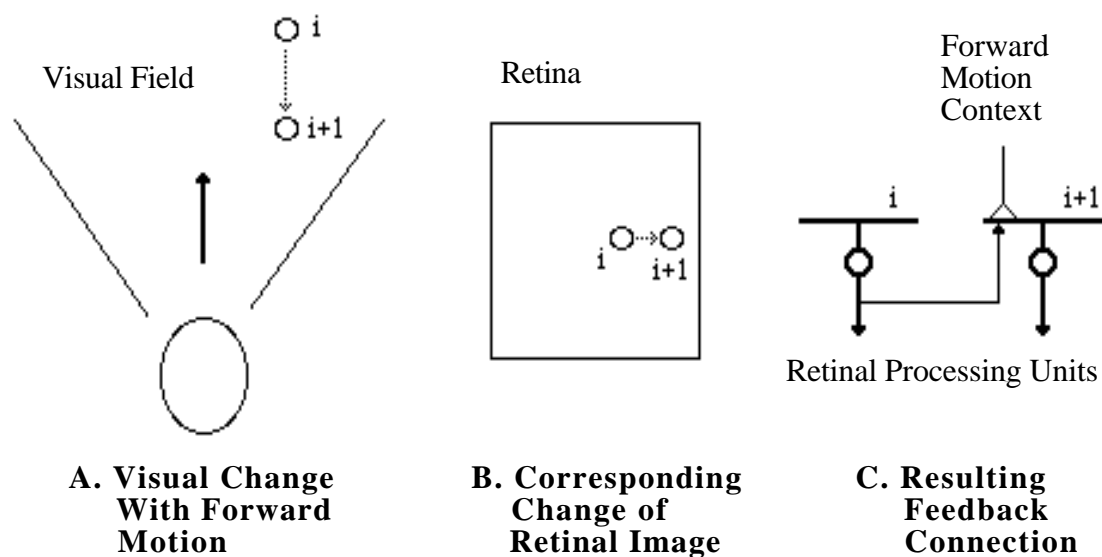


Figure 3.5: Effects of movement on retinal projection. Adapted from Mel (1986)

During 'forward motion' (A), the retinal projection of an object changes, typically as radial motion away from the center of the retina, provided the center of the retina is aligned with the direction of motion (B). Thus during learning, the connection from unit i to unit $i+1$ will be gated on during that motion context, as it predicts the future retinal state in that context (C).

away from the retinal center will be gated on, while others will be gated off. This is exactly analogous to the bioreactor emulator (Figure 2.8). The emulator is trained by learning a mapping from current state and control signal to next state.

The 'envisionment' phase of the model is pretty much the same thing (see Figure 3.6), only without the continuous external stimulation:

The second phase of [the model's] operation is the phase of *internal simulation or envisionment*, and runs concurrently with phase 1 -- but only becomes accurate after sufficient phase 1 learning. In phase 2, let us assume the array of [processors] is excited into some initial state of activation by the retina, when confronted by a novel 3-D object viewed from an arbitrary perspective. This internal visual state ... may be thought of as a "mental

image" ... By issuing a motion-context ... and temporarily inhibiting the retinal pathway, [the model] can transform (e.g. rotate, zoom, or pan) this mental image through time in an approximation to the internal state sequence that would be driven by the retina, were [it] actually moving through its environment and "seeing" the changes. (Mel, 1986)

This model has some attractive features. First, as Mel notes, it does not try to compute an explicit 3-D representation of the visual object (a typical goal of traditional computer vision). Rather, the three-dimensionality of the object is *implicitly* represented by the way its image transforms under rotation, etc. This is suggestive because it points to a way in which 3-D information can be represented on a 2-D sheet of processors, something which natural vision systems seem to be able to do. We can speculate that once trained, the model, even during 'normal' visual perception, sees the 2-D projections *as* 3-D²³ because of their implicit, inherent 'move-around-ability,' something that is not there before training simply because the model has not learned their 3-D characteristics. The converse of this is that had the model not been able to build the forward model of 3-D visual-motor transformations, it would not have 'normal' 3-D perception.

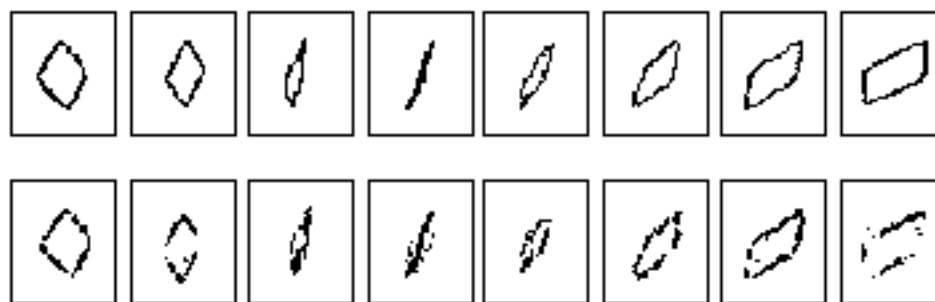


Figure 3.6 Real (top row) and imagined (bottom row) rotation. (adapted from Mel (1986))

As Mel points out, this is reminiscent of the Held and Hein (1963) result concerning the importance of motor-visual feedback in the development of normal vision.

²³ I don't think it implausible to assume that if a system can manipulate a 2-D visual representation in such a way as to preserve 3-D invariances, then it, at least implicitly, represents the object as 3-D.

In this experiment, two kittens were raised in identical²⁴ sensory environments. This was done by putting the kittens in an apparatus which maintained their location and orientation in counter-part points of a radially symmetric visual environment. (see Figure 3.7). The only difference was that one of the kittens moved itself around the environment, while the

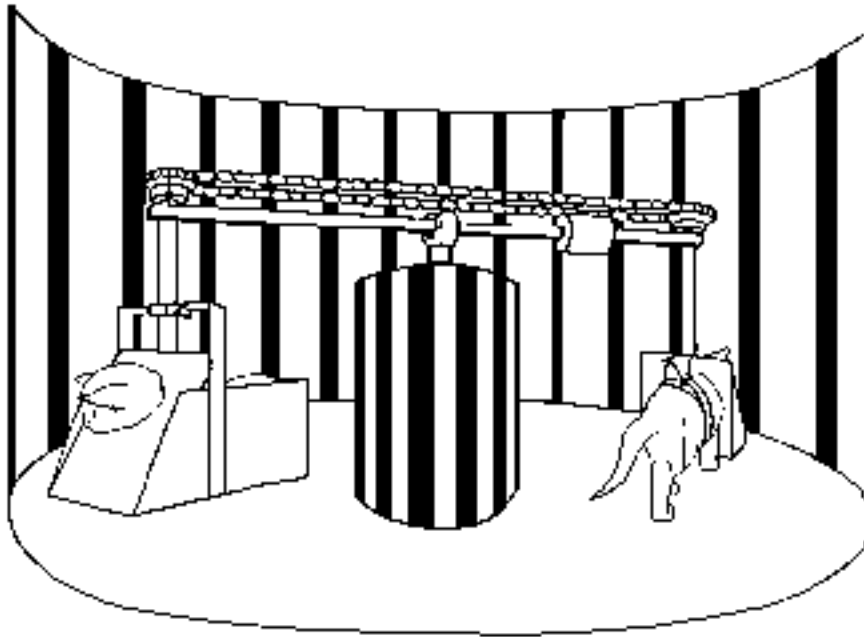


Figure 3.7: Sensorimotor feedback apparatus. Adapted from Held and Hein (1963).

other was passively moved around visually identical scenes. The result was that, even though both kittens received very similar visual information during development, only the one that moved itself around developed normal 3-D vision. This suggests that normal perception, in addition to imagery, might be dependent on emulation (this point will be explored in the next section).

Furthermore, it turns out that the 'visual imagery' of the model respects the isochrony principle, which holds that imaged events and their overt counterparts exhibit similar temporal profiles. For example, it will take the model the same time to 'mentally

²⁴ As close to identical as possible. See Held and Hein (1963) for a more detailed explication.

rotate' an object 30 degrees as it takes to move 30 degrees around that object. This is unsurprising since the model's imagery capacity was learned from its overt counterpart. That human imagery respects isochrony suggests that the mechanisms of its generation are emulators of overt counterparts as well. For example, Decety and Michel (1989) found that subjects took the same amount of time to write a text fragment and to imagine writing the same fragment. Furthermore, subjects took the same time to write fragments with their left (non-dominant) hands as when imagining writing with their left hands.

Even more suggestive is a recent result in Farah et al. (1992). The authors tested a patient on the following task before and after unilateral occipital lobectomy. She was asked to imagine familiar objects moving toward her slowly, and to determine the approximate distance at which those objects began to extend beyond the periphery of her mental 'visual field'. The finding was that the distance doubled after the lobectomy, suggesting a concomitant reduction in the width of the field of the mind's eye (as might be expected, this change was noticed only in the horizontal dimension). Upon reflection, this experiment provides startling evidence for the current hypothesis that imagery is perception emulation. The result suggests that visual imagery depends on exactly the same machinery that supports normal vision -- *all the way back at the occipital lobe!* If imagery were just somehow computed on the basis of memories, or via some translation of information in a propositional format (e.g. Pylyshyn (1984)), then it is hard to see why removal of the occipital lobe should have any effect whatsoever.

3.4 Perception and Closed-Loop Imagery

The most common view of perception seems to be that it is almost exclusively a bottom-up, data-driven process. These processes, on this view, are the sole providers of perceptual information to central mechanisms, meaning that top-down influences are minimal. This seems to contrast with the case of imagery, because in such cases there is some sort of information available, sensations or qualia of a sort, but there is no bottom information to work its way up. Imagery thus seems to be a case of top-down, or maybe sideways-in, perceptual processing. It may be illuminating to recall again the bioreactor emulator in section 2.4 (Figure 2.9). There it was explained that the emulator has two possible sources for its sensor information. The first is the sensors themselves, and the second is recurrent connections from its own outputs. We might think of the first case as a sort of pure perception, and the second as a sort of pure imagery. Though these are the two extreme possibilities, there is a continuum of cases here -- indeed a continuum along several dimensions.

First, one might not wish to ignore either the feedback connections or the sensor input, but to combine or average them in some way. Sensors can be very noisy, and the perceptual environment can make a lot of false promises as well. Were we slave to *all* the deliverances of our senses, our cognitive lives would be pretty difficult, I think. An emulator could provide exactly the sort of statistical 'soft-focus' we would want to help us clean up the noise, and to do so intelligently on the basis of significant environmental regularities. On the other hand, sometimes the anomaly is exactly what we want to see. I don't want to speculate on how these mixtures might be achieved. If that is in fact what happens, then I take it to be an empirical issue how they interact.

Another dimension of variation is one of attention. In the bioreactor emulator case, the number of feedback connections is the same as the number of sensor inputs. But what if the possible space of sensor information were greater than could be handled at any given time through direct input? For example, what if it were possible to get a *real* external reading on *only one* of the three state variables at a given time (perhaps analogous to our limitation of foveating on only a small area at a given time)? In such a case one might feed the other inputs via recurrent connections, leaving only the one to run off of external sensors. It might even be possible for the emulator to determine which of its 'modalities' it has the least confidence in at a given time,²⁵ or which it thinks is most critical to maintaining an accurate model, and focus its attention on that input by gating it to the external sensors and the others to recurrent connections. Still, the entire sensory state of the emulator, its current inputs, would be a function of both the external sensors and feedback connections.²⁶

This suggests the following picture of perception. A subject's 'perceptual world' has the capacity to operate completely closed-loop, to ignore all external input and simply run off its own outputs under the assumption that its predictions are accurate (as with imagination or dreaming). But, to varying degrees and perhaps by various means, the system can constrain the trajectory of the 'perceptual world emulator' with sensory inputs, just as one might occasionally reach out to support a child who is learning to walk, or reset a watch that has fallen out of synch. The idea is captured beautifully in the quote at the head of this chapter by Jain. On this view, perception is indeed a controlled hallucination process, the controls being provided, to a greater or lesser degree, by the senses.

²⁵ For example, the outputs might include not only an estimated sensor value, but an estimated deviation as well, and a high estimated deviation would mean a low confidence in the estimate.

²⁶ Interestingly, the view of perception put forward here has the capacity to reconcile many of the conflicting intuitions regarding the top-down vs. bottom-up views of perceptual processing. If perception is sustained as part of an emulatory *loop*, then the influences of higher centers on the lower ones enter, as it were, *at the bottom*. Seeing the process as a loop, rather than a line, allows us to have our bottom-up cake and eat it too. We will be able to agree that perceptual processing goes from the bottom up, while also agreeing that higher mechanisms can influence it substantially.

Rudolfo Llinas has recently advocated a similar position.²⁷ He argues that REM sleep and wakefulness are very similar in many important respects, the only significant difference being the capacity for external sensory stimulation to make a coherent contribution to the intrinsic activity. He argues that the thalamo-cortical loop which subserves perception and awareness is best regarded as a *closed loop* that is capable of sustained auto-stimulation. Llinas and Pare are worth quoting at length:

The thalamus is considered to be the functional and morphological gate to the forebrain. Indeed, with the exception of the olfactory system, all sensory messages reach the cerebral cortex through the thalamus. Yet, synapses established by specific thalamocortical fibers comprise a minority of cortical contacts. For example, in the primary somatosensory and visual cortices, the axons of the ventroposterior thalamic and dorsal LGN neurons account for, respectively, 28% and 20% of the synapses in layer IV and adjacent parts of layer III (where most thalamocortical axons project). Even in primary sensory cortical areas, most of the connectivity does not represent sensory input transmitted by the thalamus, but input from cortical and non-thalamic CNS nuclei. Indeed, cortico-striatal, corticocortical and corticothalamic pyramidal neurons receive, respectively, 0.3 - 0.9, 1.6 - 6.8, and 6.7 - 20% of their synapses from specific thalamocortical fibers, while less than 4% of the synaptic contacts on multipolar aspiny neurons in layer IV originate in the thalamus.

...the thalamocortical network appears to be a complex machine largely devoted to generating an internal representation of reality that may operate in the presence or absence of sensory input.

One way to think about this view of perception is that perceptual processes are subserved by mechanisms that execute *trajectory* completions, a generalization of the popular connectionist notion of *vector* completion. Normal vector completion approaches are best suited to static knowledge structures, and constitute a special, temporally flattened case of trajectory completion. Llinas' proposal is that the 'reality emulator' is a trajectory completer, or trajectory continuer, which can be more or less constrained (in the normal 'constraint satisfaction' sense of the term) by sensory input.

The possibility that our entire world is an internal emulation opens up before us.

²⁷ See Llinas and Pare (1991)).

3.5 Conclusion

Much ground has been covered in a short span in this whirlwind chapter, and I am painfully aware that there are many important issues and counter-positions that I have ignored. But my goal has been to discuss specific possibilities, rather than demonstrate necessities.

I hope that by now the notion of emulation is clear. In this chapter and the last we have discussed emulators of robot arms, bioreactors, human musculoskeletal mechanics, perceptual apparatus, and even perceptual reality itself. In each case I have tried to make the point that it is plausible to suppose that the sort of emulation envisioned actually occurs. Furthermore, within each domain discussed, there are well-respected researchers who agree. I hope that the clear application of emulation to enhance motor control makes the strategy's phylogenetic appearance plausible. Finally, I hope to have captured the reader's imagination and charity. No doubt there are plenty of gaps in the story to fuel critics' fires, but I think there is enough of interest, and sufficient potential, to justify optimism as well as a little patience.

Chapter Four: Emulation, Representation, and Learning

In the previous chapter, we saw how adopting ETM can shed light on phenomena as diverse as motor control, imagery and perception. In this chapter we will try to do the same for cognitive development. In the first section we will look briefly at what is generally agreed to be a developmental watershed, which occurs around the middle of the second year -- the capacity to entertain hypothetical events. This section will be short, because much of what needs to be said has already been addressed in the previous chapters, and Chapter Five will explore this capacity as well. The remainder of this chapter will examine some further aspects of development, specifically the Representational Redescription (RR) hypothesis.²⁸ I will show that by elaborating ETM slightly, we can account for much of the machinery of RR, to the point of, perhaps, making some finer distinctions than the RR model itself.

4.1 Dreaming, pretense, and altered control loops

In a number of works, most obviously *Play, dreams and imitation in childhood*,²⁹ Piaget argued that the end of the first developmental stage, the sensorimotor stage, was signaled by a number of semiotic capacities. Such capacities have their roots in deferred imitation, the ability to imitate some motion or action some time after it has been observed, but have their fruition in such phenomena as pretend play, dreaming, and imagination. These

²⁸Cf. Karmiloff-Smith, A. (1992).

²⁹Piaget (1962).

capacities are argued to then underwrite linguistic abilities to various degrees. What unites these capacities -- dreaming, imagination, pretend play -- is that they are *representational* in one way or another. In pretend play, the shoe box *represents* a car. In imagination and dreaming, the vehicles are internal, but are representational nonetheless.

I have of course stolen my own thunder in the previous chapter, where I argued that dreaming and imagination are supported by the internalization of the control loop, which requires an emulator. But notice that pretend play conforms to the same schema, the control loop (too formal a term in this context, but I'll stick with it) is still external, but does not follow its normal path. It is still use of a model or models, so to speak, but not internalized ones. Control loop manipulation, the ability to disengage from the normal target system and to engage something else as a surrogate, has been implicit in ETM all along, but now is a good time to make it explicit. This ability subserves not only pretend play and eventually control loop internalization, but also the use of external symbol systems of certain sorts as well.

Going into more detail here would take us too far afield (though I will return to this topic in section 5.3). I want in this brief section merely to make contact with a robust phenomenon in developmental psychology, and show how ETM sheds some light on it. Our main concern, however, is still the *nature* of, and uses of, the internal emulator, and not simply its existence.

4.2 Representational Redescription

In previous chapters emulators have been defined roughly as input output devices, and as such the manner in which input is transformed into output has not been a concern -- we must now forgo this luxury. It has become increasingly clear that in order to understand learning and development, especially in the case of humans, one cannot pay exclusive attention to performance. In a great many domains, in other words, what makes experts expert is not just

what they do, but *how they do it*. This seems paradoxical initially, because expert performance is just a type of performance, and it would seem that what separates expert from novice is some criterion of success. The reason that this is becoming a less adequate characterization of expertise is that the skills one brings to a problem to solve it will often determine the degree to which one will be successful at similar but non-identical problems, or at the same problem under different conditions. More generally, the way one solves a set of problems determines not only how one will perform on those problems, but also the dimensions along which one can alter the problem set while remaining successful. For example, the student who solves physics problems by looking up the answers in a stolen instructor's manual has a skill that will generalize well to the case where the time allowed for each problem is reduced from ten minutes to one minute, but will generalize poorly to the case where the values of some of the parameters have been changed.

I will approach these questions by leaping from the giant shoulders of Kawato, Mel, Jeannerod and Llinas in the previous chapter, to those of Annette Karmiloff-Smith. In particular, Karmiloff-Smith has articulated a theory of cognitive development³⁰ which sees crucial aspects of the mastery of some problem domain not as changes in performance, but rather as changes in the representational and computational resources used to approach the domain.

My claim is that a specifically human way to gain knowledge is for the mind to exploit internally the information that it has already stored (both innate and acquired), by redescribing its representations or, more precisely, by iteratively re-representing in different representational formats what its internal representations represent. (1992, p.15)

³⁰ Though Karmiloff-Smith presents her theory and framework as directed towards development, I think it is equally applicable to many other cases of learning, even in adults. In fact, many of the examples she uses to illustrate her theory are in fact from adult learning, such as her own experience in learning to solve Rubik's Cube (Karmiloff-Smith (1992, p. 16-17)).

My plan is, first, to briefly outline the RR Model, and highlight certain features of it. Then, we will return to the robot arm emulator from Chapter Two, and discuss three different ways in which such an emulator might do its job. Then, with these distinctions in mind we will return to the RR Model, and to the phenomena it addresses, and show that ETM can go some way towards shedding light on the mechanisms of RR. One caveat before I continue, however. I do not want to argue that all the interesting developmental phenomena will be addressed from within the ETM framework. I think that there are clearly aspects of learning and development that have little or nothing to do with emulation (for example, the operation of inverse mappings). Rather, what I hope to do is to show how ETM, suitably refined, provides us with an apparatus for approaching aspects of learning and developmental change.

The RR Model posits four distinct types of representational³¹ format that may underwrite the capacity to succeed in some domain. The first of these Karmiloff-Smith calls Level-I representations (for Implicit). This level is characterized as follows:

- Information is coded in procedural form.
- The procedure-like encodings are sequentially specified.
- New representations are independently stored.
- Level-I representations are bracketed, and hence no intra-domain or inter-domain representational links can yet be formed.

The idea is that on any given occasion where some successful action is performed, this action sequence as a whole is stored (this idea will be elaborated shortly). This sequence, as a whole, is then available to be called upon when the same situation arises again. But, information as to *why* this sequence works, or as to the possibility that the sequence can be analyzed in terms of subsequences that solve distinct aspects of the problem, is unavailable. Such information remains implicit in the bracketed sequence as a

³¹ Throughout this chapter I will follow Karmiloff-Smith in using the term 'representational' to cover what is really a sort of composite of representational format and computational procedures.

whole. Furthermore, it is possible that one can achieve 'behavioral mastery' of the domain in question by storing a sufficient number of such experiences, such that for any situation that is likely to occur, one has a stored sequence that leads to success in that situation.

Level-E1 representations are the next level 'up'.³² An interesting aspect of the RR theory, which will be elaborated later in this chapter, is that it claims that the shift from Level-I to Level-E1 occurs only after 'behavioral mastery.' That is, it is only after some domain can be reliably negotiated by means of the Level-I representations that Level-E1 representations emerge. Level-E1 representations are not, therefore, constructed in order to increase performance or because of some current shortcomings. (Later I will suggest a theoretically motivated reason for why this might be so.) This shift involves several changes from Level-I format. In fact, as I will argue, some of the changes Karmiloff-Smith alludes to seem sufficiently distinct that it is unclear why she sees the transition from Level-I to Level-E1 as a single step, or as all being manifestations of a single format change.³³

Some aspects of this shift are:

- a) A recognition of subsequences which solve some aspect of the problem, and which are repeated within a single sequence, or are common to many different sequences,
- b) An ability to predict the outcome of parts of some subsequences,
- c) A compression of information, and concomitant loss of detail,
- d) An ability to break free from the sequential and temporal constraints of the bracketed sequences,
- e) An ability to support counterfactual conditions and 'pretend play.'

Features (b) and (e) follow directly from the use of emulators. Prediction, as in (b), was the main motivation for the use of emulation in Chapters 2 and 3, and support of

³² Level-E2 and Level-E3 are given much less attention in Karmiloff-Smith's work, but basically they involve the redescription of Level-E1 knowledge into forms that are more readily interpretable and usable by the cognitive system as a whole -- translation into a 'system-wide code'. This sort of redescription is taken to be a prerequisite for verbal description of the representational resources brought to bear on the domain.

³³ I made this same complaint in Grush (1995b).

counterfactual conditions was discussed in section 2.4. Pretend play was discussed earlier in this chapter. Features (a) and (c) are multifaceted, but involve some notion of extracting information in a more compact form. These are the features that will be the focus of this chapter. Feature (d) is quite fascinating, but I am unable to address it at this time.³⁴

As an example of the difference between Level-I and Level-E1, consider the problem of programming a VCR. One way to proceed is to have a list of button-pressing sequences that one can employ to achieve certain specific goals. Thus, to tape channel 3 from 6:00 pm to 7:00 pm, press the ON button and then press sequence A. Such a facility with VCRs may result in behavioral mastery, provided one's list of sequences is long enough. But (as is the case with my parents, for example) should anything happen during the sequence, such as a missed button press, one must start over from the beginning, i.e., turn the VCR off and start again. A more sophisticated approach involves the understanding of the VCR as a system with various states (record mode, programming mode, play mode) and a mastery of techniques for navigating between these states and for making corrections within them. Such a skill allows one to backtrack, to correct mistakes on the fly, to operate the system intelligently. One is no longer just pushing button X and then button Y, but is rather putting the VCR into record mode and setting a speed. One represents, in other words, the other side of the control loop not just as a black box, but as a system with a certain amount of internal structure.

³⁴ In fact, I think that there is a process of schematization at work in (c) and (d) which I will not be able to address directly.

4.3 Ways of emulator making

4.3.1 Level-I and lookup tables

We now return to the robot arm of Chapter Two. I want to take as an example one particular arm motion (the details don't matter, just think of your favorite arm motion and keep it in mind), and to examine, in this and the following two subsections, three different ways in which an emulator might perform the appropriate input/output mapping. The arm, we suppose, starts in an initial state S_i (a specification of the relevant dynamic variables), and the operator produces a sequence of effector commands, $A_1 - A_n$, and as a result the arm ends up in state S_f . The emulator's job, recall, is to determine the final state³⁵ as a function of the initial state and the action sequences. We might suppose that this could be computed by simply recalling from memory a previous experience where the arm was in the same initial state, and where the same sequence of actions was performed, and then reading off what the final state was on that previous occasion, and using *that* as the prediction in the present case.

This ability assumes that there is a reasonably robust store of remembered experiences. Such a system would construct this lookup table of past experience by monitoring the operation of the arm, operations which are perhaps initially random. For every action of or on that system, the system creates a new 'entry' consisting of

- a) the initial state of the target system
- b) the actions performed, and
- c) the final state of the target system.

³⁵ This is a bit inexact. The emulator may be required to determine just the final state, or perhaps to provide a continuous supply of information as to the state changes as the commands are issued. This second task may or may not be possible, depending on the way the emulator is implemented.

This store of memories, then, would be a list of the form

- (1) [S₂₃, A₁₁, A₂₃₈, ..., A₇₁ : S₁₀₉]
 [S₄₅, A₄₁, A₈, ..., A₃₃ : S₁]
 [S₈₃, A₁₈, A₁₂₄, ..., A₆₁ : S₂₉]
 [S₁₁, A₅₄, A₉₈, ..., A₁₇ : S₈₂]
 etc..

where S_x is some state, and A_x is some action. What the table then does in order to act as an emulator is, upon receipt of the initial state specification and efferent copy information consisting of the command sequence, it searches for a match in its list of stored input/output pairs, and if one is found, it produces the 'final state' specification on that entry as output.³⁶ Such a list can, of course, not only perform adequately as a forward model, as just described (assuming, of course, enough entries and stable target domain dynamics), but it can equally well act as an inverse model (which seems to be more along the lines of what Karmiloff-Smith has in mind for a Level-I representation). If the target system is currently in S_x and one wants it to be in S_y, simply look for an entry in the table whose initial state specification is S_x and whose final state specification is S_y, and perform the action sequence listed on that entry. Or to put the point in Karmiloff-Smith-friendly language, for nearly any goal, there is a stored action sequence that will reliably result in the achievement of that goal.

4.3.2 Parametric Functions

Another way in which one could implement an emulator (and inverse mapping) is by means of a parametric function.³⁷ This would have the form of some mathematical

³⁶ A proposal similar to this, for learning the forward model of the dynamics of a robot arm, can be found in Atekeson and Reinkensmeyer (1990).

³⁷ Many connectionist networks, including backprop nets, are parametric functions. The parameters in this case are the weights.

formula with variables for the initial state specification and for the action sequence, such that when solved, it specifies the values of the final state specification.³⁸ In other words

$$(2) \quad \{f: (S_i, A_1, A_2, \dots, A_n, P_1, P_2, \dots, P_m) \rightarrow (S_f)\}.$$

One could do the same thing with the inverse mapping, and use a function such as

$$(3) \quad \{g: (S_i, S_f, P_{m+1}, P_{m+2}, \dots, P_n) \rightarrow (A_1, A_2, \dots, A_n)\},$$

where P_x is some parameter. Modulo the initial state specification, g is the inverse of f .

Notice that there is a straightforward sense in which both (2) and (3) are implicit in (1).

There is also a straightforward sense in which (2) and (3) represent 'compressed' versions of (1). Thus parametric functions seem to offer some of the features of Level-E1

representations. These two points notwithstanding, I think that such parametric functions are a poor choice for Level-E1 representations. Though (2) and (3) are implicit in (1), and this gives the impression that (2) and (3) are making something explicit, what they are making explicit is some mathematical function which may or (more likely) may not shed any light on the operation of the target system.³⁹ They both provide information that is still inflexible and bracketed. The action sequence one gets by using (3) is just as unanalyzed and compositionally opaque as the one obtained by using (1).

4.3.3 Dynamic Analog Models

Let us turn for a moment to the arm itself. The arm's state changes according to the laws of physics. The state variables, such as angle, position, angular acceleration, etc., are all coupled dynamically by the laws that govern the behavior of physical objects. We might

³⁸ In essence, this is what a vanilla three-layer backpropagation network would do, if it had an input node for each value that needed to be specified (for both the initial state and for the action sequence), and had the same number of output nodes as values in the final state specification.

³⁹ This approach has led to the use of such devices as principal component analysis and cluster analysis in an attempt to see what, if anything, of interest the parametric function is representing.

plot the values of some of these variables as in Figure 4.1. In this figure, (a) and (b) represent the shoulder and elbow angles, respectively, during the course of the movement. The angular velocity and acceleration of the elbow joint is given in (c) and (d), respectively, while (e) plots the angular inertia of the arm, which increases as the elbow joint straightens. (f) and (g) plot the agonist and antagonist torques to the shoulder joint, and (h) represents their sum, the total torque applied to that joint.

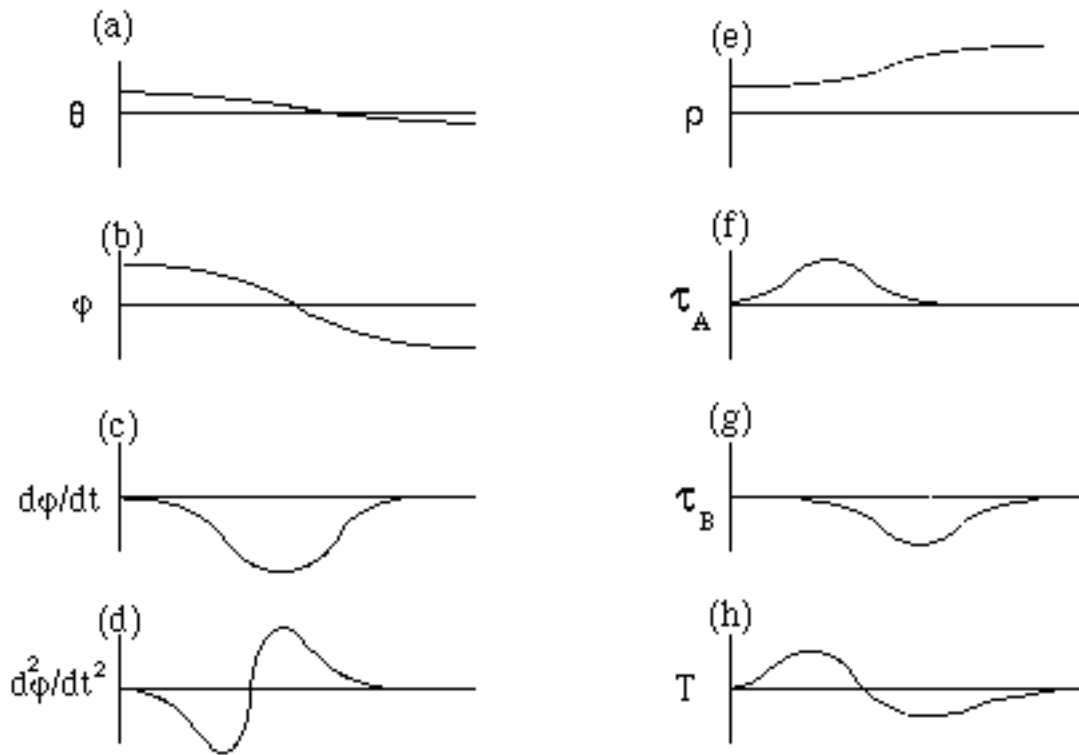


Figure 4.1: Plot of some dynamic variables during a typical arm motion.

These are just a few of the many interdependent physical parameters that conspire to determine the evolution of the dynamic system. Looking at the target system's evolution as governed by the interaction of discrete yet richly interdependent physical parameters, many of which are highly abstract, suggests a further manner in which we might implement an

emulator of this system. For each physical parameter that plays a role in the evolution of the system, assign a processing unit, such as a neuron (or a pool of neurons), whose output at any given instant is proportional to (or perhaps a monotonic function of) the value of the physical parameter which it represents. Interconnect the units in such a way as to mirror the relationships between the physical parameters (for example, the unit representing elbow angle is connected to the unit representing total angular inertia in such a way as to mirror the effect of increased elbow angle on angular inertia, i.e. an excitatory connection). Thus, when some of the parameters are specified via efferent copy information, the entire ensemble is set in motion, the units' activities and interactions, under the physical interpretations used above, exactly mirroring, in real time, the real physical parameters governing the real arm in motion (and thus the plots in Figure 4.1 would equally well be time plots of the spike frequency of such units) -- an intricate interpretive dance between perhaps thousands of neural partners, choreographed by experience, and stepping in time with the dynamic evolution of the external system.

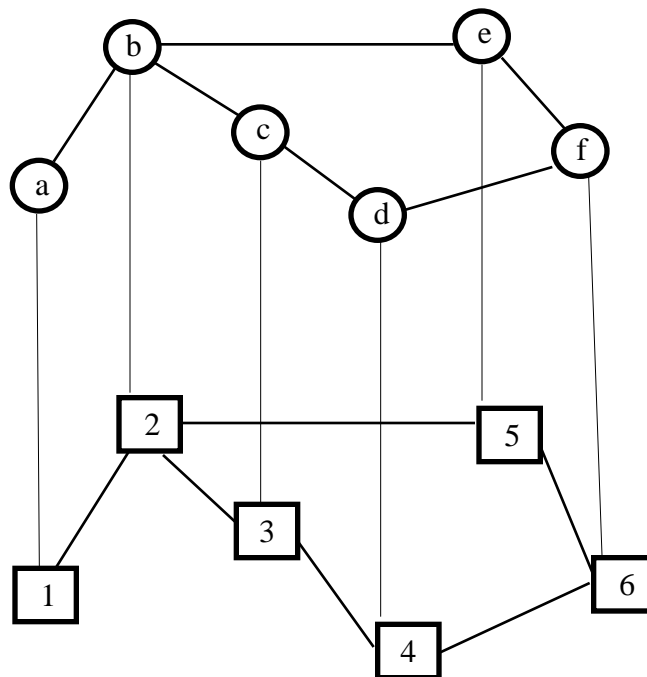


Figure 4.2: Schematic mapping between physical parameters and analog model.

Such an emulator is much more than just an accurate mimic of input/output operation, but is in effect a *theory of the target domain*.⁴⁰ While the lookup table, the parametric mapping and the dynamic analog model all allow prediction of the target system, only the latter affords an *understanding* of its operation.⁴¹ Consider Figure 4.2. The circles in the upper part of the diagram (a - f) represent emulator articulants (instantiated perhaps as neurons, or groups of neurons). The lines interconnecting them represent the mutual influences that these neurons have on each other. These influences could be more complicated than a mere scalar connection strength. The squares in the lower part (1 - 6) represent physical parameters that govern the dynamics of the target system. These might be (in the case of an arm) hand velocity, elbow torque, mass of the forearm, etc. The lines interconnecting the squares represent the mutual influence that these parameters have on each other. For example, changing the elbow torque has an effect on hand acceleration. These lines, some of them at least, will correspond to dynamic couplings of various sorts. The vertical lines are interpretation lines, and are meant to indicate the fact that the units in the analog model *represent* parameters of the target system. As mentioned in the previous chapter, neurons have been identified in the magnocellular red nucleus of the cat whose spike frequency profile closely matches physical parameters of the cat's forelimb in motion.

⁴⁰ Of course, the theory thus implemented may be false, and the understanding it fosters a sort of false understanding, but this is not to be confused with not having a theory at all.

⁴¹ We are now in a position to make contact with some of Paul Churchland's views on the relationship between neural computation and scientific explanation. Churchland characterizes a theory as a point in weight space, and levels effective criticism against certain 'higher level' characterizations, such as cluster analyses, or activation space partitions, as being adequate descriptions of the dynamics of neurally implemented theories (Churchland, 1989). Though I don't have the time to address this adequately, on the present account, Churchland's position is either incomplete, or invokes a different notion of 'theory' (and I would claim further that the notion as I am using it is closer to common parlance). On the present account, a (causal) theory is a dynamic analog emulator whose components' interactions are taken to mimic the laws and properties which govern some target domain. Not just any point in a weight space will fill these criteria. Churchland may well agree, claiming that, while true that a theory is a point in weight space, he didn't mean to imply that any old point in weight space could be a theory.

If my interpretation is correct, then these would be an example of the type of analog emulator here envisioned.

4.4 What can ETM say about development?

The complaint might be raised at this point that I have said a lot about representation and implementation, and very little about learning or development. In the subsections that follow, I will briefly indicate the main ways in which ETM, as articulated, can address issues in learning and development.

4.4.1 The block balancing task

One of the tasks examined by Karmiloff-Smith is the block balancing task. Consider the blocks in Figure 4.3, all of which are balanced on a narrow support. Block A has an embedded lead slug, which moves its center of mass far from its geometric center. The task is to balance the blocks on the support, an activity that children spontaneously participate in. The results Karmiloff-Smith reports are that young children of 4 - 5 years do this task quite well. They place a block on the support, and use feedback concerning the block's direction of fall to reposition it.

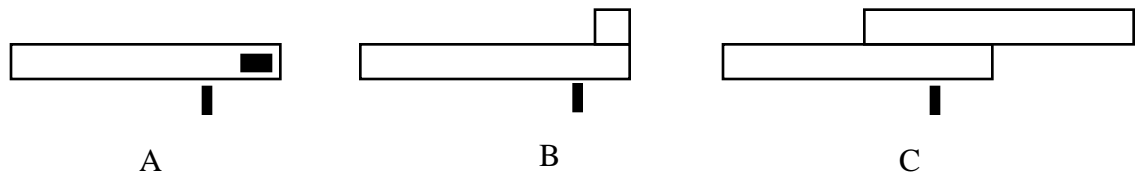


Figure 4.3: Blocks from the block balancing task.

By contrast, 6-year-olds immediately place blocks on the support at their geometric center, regardless of their apparent center of mass. In fact, as Karmiloff-Smith reports, the

geometric-center theory is so strong that even after the block falls, "they put it right back at the geometric center - but very much more gently!"⁴² Older children successfully use both geometric and center-of-mass information available proprioceptively to balance all blocks fairly well, placing them near the appropriate spot on the first attempt. Such children then appear to have an internalized savvy concerning the major dynamic factors that determine where blocks will balance.

Piaget reports similar results concerning children's ability to understand factors (thickness, length, material, etc.) responsible for determining the flexibility of rods.

When asked to furnish proof [of why a given factor makes a difference - RG], subjects of nine or ten will choose a long thin rod and a short thick one to demonstrate the role of length, because in this way, as a boy of nine and a half told us, "you can see the difference better"! From eleven or twelve onward..., subjects, after a little groping, make a list of factors by way of hypothesis, and then study them one by one... That is, they vary each factor alone, keeping all the other factors constant. For example, they choose two rods of the same width, the same cross section..., and the same substance, varying only length. This method... is all the more remarkable in that none of the subjects we interviewed had received instruction in this method at school. (Piaget and Inhelder 1969)

The crucial point is that there seems to be an ability, manifested in development and perhaps in learning generally, which allows not only for behavioral mastery of some domain of activity, but for the internal representation of separable factors that determine the operation of the domain in question. This point bears emphasis, especially in light of a current trend in theoretical cognitive science that attempts to minimize the role of internal representation and highlight the role that the environment (and agent/environment interaction) plays in successful activity.⁴³ While much activity that has previously been taken to require internal representations of some sort may in fact be amenable to this style of analysis, it seems clear that there are cases that cannot be handled in this way. An

⁴²Karmiloff-Smith (1992) p. 86.

⁴³Cf. van Gelder (1991), Brooks (1991).

interactional account might be a plausible analysis of what the 4-year-olds do in Karmiloff-Smith's block balancing task, but it isn't clear that Karmiloff-Smith's nine-year-olds, or Piaget's twelve-year-olds, are operating in this manner.

4.4.2 Control loop internalization as a developmental watershed

One immediate application of ETM to learning and development, which does not depend on the details of the previous sections, is that it shows exactly what it means to have an internalized control loop. One of the best confirmed and most widely attested 'stages' in development is the onset of imagination at around 18 months (even Karmiloff-Smith admits that this may be a domain-general watershed (1992, p.167)). Imagination is clearly (as argued in the previous chapter) supported by an internalized control loop, and pretend play is supported by an altered control loop (this will be addressed in more detail in the next chapter). Given the importance that imagination and pretense play in the development of other capacities (such as initial theories of mind -- see Chapter Five), its characterization within the ETM framework is crucial.

4.4.3 The role of behavioral mastery

As mentioned in section 4.1 above, the RR Model posits that the more flexible levels of representation emerge only after an initial period of behavioral mastery. Within the ETM framework, and assuming temporarily an equivalence between Level-I and lookup tables, and an equivalence between Level-E1 and dynamic parametric models, we might be able to address the purpose of behavioral mastery. The idea is that lookup tables are relatively quickly learned, in view of the fact that the only learning necessary is some sort of *memory*. Furthermore, it might be the case that the intricacies of dynamic parametric models require so much training (in order to function well at all), that most of their training must be done not just from experience with the real world, but from internal training on the

Level-I (lookup table) representations.⁴⁴ In a sense, then, the model would be a theory not only of the target domain, but (what amounts to the same thing) a theory of the knowledge implicit in the Level-I representations themselves. This seems to be more along the lines of what Karmiloff-Smith has in mind when she describes representational redescription as "abstract[ing] knowledge the child has already gained from interacting with the environment." (1992, p.78)

4.4.4 Redundancy and true flexibility

One of the claims that the connectionist orthodoxy pushes is that connectionist models of domain knowledge are flexible in a way that classical models are not. It has been pointed out, however, that flexibility can be assessed along many dimensions, and that there are ways in which connectionist models are quite inflexible where classical models are flexible. One moral of the preceding discussion is, I think, that true flexibility lies not in having a 'flexible' implementation of domain knowledge, but in being able to flexibly access the right implementation, for the task at hand, from a range of alternatives. Thus, the expert will have at her disposal a vast amount of lookup table experience of the Level-I sort, a theory (an internal dynamic parametric emulator) of the task domain, and perhaps a few sorts of inverse models as well (i.e. a short list of well-practiced, 'chunked' routines, a larger experience of past attempts at control that failed, etc.), and no doubt a host of other tricks and well-worn techniques.

For example, consider a ten-year-old confronted with a new version of the block balancing task conducted in a room with a strong magnetic field, a version in which some of the blocks are weighted with iron and others with lead. The child's internalized model of the relation between center of mass and geometric center will no longer work well, but one could imagine that after a few failures, the child reverts to the much older strategy of

⁴⁴ Andy Clark (1994) raises a similar possibility, of two networks, one a quick-learning lookup-table and the second, a slower gradient descent model which learns from the first.

placing the block on the support and then adjusting its position on the basis of which way it starts to fall. The point is that under the more general rubric of forward models (and even inverse models), it is possible to discern a number of representational formats, some of which may depend on the prior acquisition of others, which bring different capacities and flexibilities to the task domain.

4.5 Parameters, distributed representations, and semi-locality

It is now time to make explicit something which has been implicit in the preceding discussion. In the case of a dynamic analog model, the individual parameters are represented relatively locally, in the sense that the resources used to represent the value of one of the parameters are not also used to represent the value of any of the others at the same time. This of course does not imply that each parameter is represented only by a single unit or neuron, nor does it imply that, if represented by a group of units, they all have the same value.⁴⁵ It also does not imply that resources are dedicated to representing a single entity and parameter for any length of time beyond the bit of reasoning or computation in question. What it does imply, however, is that the hardware that acts as the vehicle for the representation of one of the values *at a given time* does not also serve as the vehicle for representing any of the other parameters *at that time*. This is a departure from established connectionist orthodoxy, which has long touted the virtues of distributed representations.

But in this case at least, there seem to be overwhelming virtues to localist coding. First, the local coding of the value of a parameter makes that parameter, and its value, accessible in a way that is impossible for distributed coding schemes. If, while the emulator is running, the operator (or motor cortex, or whatever) wishes to change the value of one

⁴⁵ There might, for instance, be a large number of units whose 'vector average' represents the value.

of the parameters, it can access that parameter in isolation. If, for example, I want to increase the angular acceleration of the shoulder joint, this can be done by sending a given signal to a given group of neurons. In truly distributed coding, trying to change one of the 'parameters' is very tricky. This is not so much because the resources overlap. Otherwise it would simply be a matter of sending a selective signal to the entire ensemble. What is tricky is that in a truly distributed representational scheme, sending such a signal to the entire ensemble will have different effects on the targeted parameter depending on the values of the other parameters that are represented with the same resources (in much the same way that what needs to be done to add 1 to a binary representation of some number depends on what number the entire ensemble is currently representing).

This ability to selectively address a single parameter lies, I think, at the heart of what Karmiloff-Smith means by increased explicitization, and also seems to be the core of any sort of translation into a 'system-wide code.' The fact that, as I argued, distributed representational schemes are ill-suited for such selectivity may explain why many connectionist models, as Karmiloff-Smith argues, seem incapable of the sort of development her RR Model posits.

Nonetheless, there is a sense in which what a certain articulant represents depends crucially on what other articulants it is coupled to, and the nature of that coupling. It is only because a given representation is connected in the right way to shoulder torque and elbow angle that it can represent shoulder angular acceleration, for example. I thus want to say that articulants are *semi-local*. They are local in that, as argued above, they can be independently targeted by cognitive operations. The 'semi' is meant to be a recognition of the fact that articulants are articulated from a rich representational matrix, with the consequence that altering the value of one of the articulants will alter others. When I imagine strongly contracting my forearm flexors, the sensation of my fingers digging into

my palm is spontaneously generated (unless I imagine a tennis ball in my hand, or imagine tensing my forearm extensors at the same time).

But this interdependence has limits. I can indulge in the proprioceptive imaginings above without specifying the position of my shoulder, or my posture, for example. In fact I have just done so. The character of what I have meant so far by semi-local is captured here. The musculoskeletal emulator has an articulant for the force applied to my forearm flexors. This variable can be selectively targeted, in other words. However, this articulant will invoke certain others -- most immediately the tension in the forearm extensors and the tactile sensations evoked by a clenched fist, less saliently the angle of my elbow, and less saliently still my posture and quadricep tension. There is a clear gradation or fish-eye effect manifested here. We will return to this characteristic of articulants in Chapter Six, where we will see that their semi-locality leaves clear footprints in the structure of language.

4.6 Concluding remarks

This chapter has not really grappled in detail with any particular area of learning or development.⁴⁶ Rather, I have focused on a theory, the RR model, which has been applied to a number of specific domains, and I have compared the machinery of emulation to some of the posits of that theory. I thus am assuming that the RR Model, and the reasoning behind it, are well motivated, and that the compatibility between the RR Model and ETM as articulated in this chapter licenses ETM to enjoy the fruits of the detailed analyses given by Karmiloff-Smith. There are benefits for ETM, as well as for the RR Model, from this comparison. ETM allows us to explain, and in some cases sharpen, some of the processes posited by the RR Model. For example, we have seen that the RR model collapses at least two separate processes (control loop internalization and emulator articulation) together. On

⁴⁶ Though Chapter Five, 'The Mind-Body Solution', will apply the apparatus to a specific area of development.

the other hand, the RR Model convinces us that when applying ETM to human, or suitably sophisticated, cognition, we cannot afford to be concerned exclusively with the success of an i/o mapping, because the details of how that mapping is implemented make a difference.

Chapter Five: The Mind-Body Solution

Infants and young children face a number of challenges in their attempts to understand the world. The challenges don't stop when one has learned more or less to control one's body, and when one has some idea of how physical objects typically work. There remains a large class of entities, namely animals and people, whose conduct is quite dissimilar to that of other entities such as chairs and balls. These entities move themselves around without obvious influence from the outside. Many of them make noises at each other, and unlike spoons and blocks, these noises can often cause them to engage in certain repeatable activities. Though there are many distinctions at work here, living vs. non-living, animal vs. non-animal, self-initiated motion vs. no self-initiated motion, etc., I want to focus on the distinction between entities that can usefully be given psychological descriptions and those that cannot. People and animals seem to engage in goal-directed behavior, they seem to know things about the world, and they perhaps even represent the world to themselves.

There are a number of ways to approach these issues. In this rather brief and relatively straightforward chapter, I will approach them from the point of view of the child attempting to understand what makes people (and animals) tick. This is one reason I made separate mention in the last paragraph of the possibilities that people know things about the world, and that they represent the world to themselves. These possibilities are at least logically distinct, and I will in fact argue (following Perner (1988)) that children go through a stage where they represent people as knowers, but not as representers. And the crucial leap from representing representers as knowers to representing them as representers is the silent revolution which overthrows many of the puzzles of their behavior. The distinction

between mind and body is not, as generations of philosophers have supposed, a *problem* which needs to be solved. Rather it is an elegant *solution* to a family of problems in the explanation of intelligent, goal-directed behavior.⁴⁷

In the next section I will quickly review some major issues associated with what is typically called the development of a theory of mind. Anyone who has read this far in my project will not be surprised to hear that my survey will be quite schematic. I don't want to review the literature, but to provide an orientation for the arguments that follow by briefly discussing the work of two major researchers in the field, Henry Wellman and Alan Leslie.⁴⁸ Section 5.2 will focus on the model proposed by Josef Perner, which seems to me to be the most clear and correct. Section 5.3 will return to the general discussion of the phenomena, and of the theories, and explore some of their shortcomings in the light provided by Perner. The final section will argue that Perner's model can be seen as a natural extension of ETM.

5.1 How the Folk Get Their Psychology

The best entry point to this discussion is the false belief task.⁴⁹ A standard version of this experiment has the child watch a character, such as a puppet named Maxi. Maxi looks on as a bar of chocolate is hidden in some location, e.g. the kitchen. Maxi then leaves, and while out, the chocolate is moved to a new location, the living room perhaps. Upon Maxi's return, the child is asked "Where will Maxi look for the chocolate?" Children older than four typically succeed by correctly answering that Maxi will look in the kitchen.

⁴⁷Of course there are many aspects of the mind-body problem besides the problems of representation and content. I am inclined to simply follow Churchland (1979) and maintain that all such aspects are ultimately matters of internal theory, and can be treated in ways analogous to the manner in which I'm treating representational states.

⁴⁸Those interested in reviews might consult Wellman (1990), the Introduction to Astington et al. (1988), and Special Issue of the *British Journal of Developmental Psychology* 1991, vol. 9.

⁴⁹ Cf. Wimmer and Perner (1983), Perner, Leekham, and Wimmer (1987).

The children know that Maxi wants the chocolate, and they also know that Maxi believes that the chocolate is in the kitchen, even though it is really elsewhere. Younger children typically fail, answering that Maxi will look in the living room. They thus predict behavior based on the desires of the agent, together with their own knowledge of how the world really is.

There are many questions one could raise about this experiment. Was it properly controlled? Does it presuppose too much command of language to properly get at the theory of mind issue? Would success on this task show that children have a developed folk psychology? These details will not concern us. What will concern us is the point of the experiment. Certainly by the age of six, children know that the actions of people are often governed by their desires, in conjunction with the way *they think* the world is. At some much earlier age, they seem to think that behavior is governed by desires in conjunction with the way the world (*really*) is. And perhaps at a still earlier age, children view most of the behavior of people as chaotic and mysterious. Whatever the details, there is a transition in the way children explain the behavior of others. This change, and the mechanism(s) underlying it, will be our concern.

Let us focus on two stages -- for simplicity Stage One and Stage Two. Stage One will be the age just before children can succeed at the false belief task. At this stage, children are nonetheless doing some impressive stuff. They know that agents have goals and act to fulfill these goals. For example, if told that Sam wants to take his rabbit to school, and that the rabbit might be in either of two locations, children will predict that Sam will look in the other location for the rabbit if it is not found in the first location (even if Sam finds something really snazzy in the first location).⁵⁰ It is therefore not the case that children make no distinction between agents and inanimate objects. Stage Two will be the stage at which children reliably pass the false belief task. Passing this task involves not just

⁵⁰Cf. Wellman (1990, chapter 8).

knowing what the agent wants, but also a sensitivity to the agent's *representation* of the world. In particular, it involves the capacity to realize that an agent might be wrong about the world, and that this erroneous representation will lead to actions very different from actions based on the way the child thinks the world is. Though this may not be a full-fledged folk psychology, it is certainly a step up from Stage One.

Leslie: Metarepresentations

An admitted adherent to Fodor's Language of Thought hypothesis, Alan Leslie (1988) attempts to explain the difference between Stage One and Stage Two (recall that this is *my* terminology) as an enhancement of the logical manipulations available for operating on mentalese sentences. These logical manipulations, or metarepresentations, have their first appearance (or at least, they first manifest themselves) in pretend play -- where, for example, the child picks up a banana and pretends it is a telephone, or helps her father give a bath to a teddy bear in an imaginary tub. The logically havoc-wreaking thought that the banana *is* a telephone, or that the *empty* bucket is *full* of water, is rendered systemically harmless by being 'decoupled' from its normal representational and computational function via embedding in a formula of the form: **agent - informational relation - 'expression.'** The informational relations available at this point will include things like 'pretend,' 'think,' 'dream,' etc. Expressions embedded will be of the form 'the telephone is a banana,' 'the empty bucket is full of water,' and the like.

One might think that the apparatus of metarepresentation has everything needed for understanding belief, and derivatively false belief. All one would need to do is to add 'believes' as a possible informational relation, and one could form expressions like: **father - believes - 'the empty bucket is full.'** In effect this is more or less true.⁵¹ Leslie

⁵¹Of course, with the entire expression decoupled and in quotes, the child is attributing a contradictory thought to her father. This doesn't seem like a promising way to go about understanding belief, but I won't pursue this here.

argues, however, that the transition to Stage Two requires the development of a causal theory of knowledge and action. This theory maintains that agents gain information about the world from being appropriately placed in a causal/perceptual pathway. On the other side, mental states can cause overt behaviors. It is only after one appreciates that mental states can be caused by perceptual states, and that mental states can cause actions, that one is in a position to have a use for the informational relation 'believes.' As long as the child thinks everyone is basically omniscient, the only use for metarepresentations is to understand pretense. With the causal theory of knowledge at her disposal, the child must keep track of what so-and-so knows. The antecedently available metarepresentational format is easily adopted for this use.

I will have more to say about Leslie's account in section 5.3.

Wellman: Simple Desire Psychology vs. Belief-Desire Psychology

Wellman (1988, 1990) argues that children under three possess what he calls a 'simple desire psychology,' and that it is only later that they use anything that could be called a belief-desire psychology. (The scope of 'simple' in 'simple desire psychology' is narrow: it is the *desires* that are being characterized as simple.) Wellman claims that there is a coherent notion of desire that is non-representational. This of course conflicts with the full-blooded philosophical notion of desires as similar to beliefs, according to which desires are attitudes taken to propositions, canonically expressed as complementized clauses (e.g. 'Rick desires that he win lotto'). In this *simplified* sense, however, desires are understood to be putative drives towards objects or situations. They would be states of the agent, but not representational states, in much the same way that the mass of my car is a property which, via familiar laws, keeps it in contact with the earth (most of the time), or at least is responsible, in a sense, for the forces that drive it toward the earth, though it does not *represent* the earth in any obvious way.

The transition to Stage Two, for Wellman, involves the construction of a new and better *psychological theory*, a belief-desire psychology, which displaces simple desire psychology. This belief-desire psychology is similar in many ways to folk psychology: it posits that there are internal states that are representational in nature (unlike the simple desires), and that these internal states constitute the agent's knowledge of the world. There are two points I want to emphasize. First, this account is quite different from Leslie's. The transition to belief-desire psychology is for Wellman a fundamental theoretical revolution in the child's understanding of the representational basis of intelligent behavior. For Leslie the transition involves no alteration in the understanding of agents' representational *capacities* at all (such capacities are part of the genetically provided language of thought), but simply the addition of a causal theory of knowledge. His analysis of pretend play is meant to show that the young child in fact has all the apparatus necessary for belief-desire psychology.

The second point is that it could legitimately be argued against Wellman that he has failed to provide any insight into the mechanisms underlying this transition, apart from merely labeling them as 'simple desire psychology' and 'belief-desire psychology'. Whatever the shortcomings of Leslie's account, he has at least explicated a mechanism. Perner's proposal, which will be the subject of the next section, offers the best of both worlds -- he correctly recognizes that the transition to Stage Two involves a fundamental change in the child's theory of human behavior, unlike Leslie. But like Leslie and unlike Wellman, he provides some insight into what this fundamental change amounts to.

5.2 Josef Perner: Modeling Models

Perner (1988) provides a nice synopsis of his theory. Having explained that he will be adopting a sort of model-theoretic semantics, he continues:

This exposition of models is then used to characterize children's intellectual progress in terms of *three levels of semantic awareness*. They proceed from the level of (1) *presentation*, where the child merely *has a mental model* of perceived reality, to the level of (2) *re-presentation*, where they can construct and *use mental models* to think about hypothetical situations. The final level of (3) *meta-representation*, is reached when the child becomes capable of not just using but *modeling mental models*.

At the first level, according to Perner, the model *constitutes* reality for the child, as opposed to representing it. Crucially, the child *has* a model of reality which is constantly updated by the senses and other sources of information (it is just not conceived of as a model). This model is *transparent* for the child. Like contact lenses, the model is not recognized as an entity which is responsible for the knowledge the child has, but is rather looked *through*.

The second level involves a number of abilities. First, the child must be able to remove the model from its normal function as an active on-line source of knowledge about the world. Second, she must be able to keep track of more than one model at once. At a minimum, she must be able to tell that a given model, which has been decoupled, is no longer to be taken for reality, but represents some hypothetical state of affairs, and she must be able to keep this model in mind while comparing it to the *reality* model. It seems reasonable to suppose that the child maintains a number of models, some used for reasoning, some used as buffers for the incorporation of evidence from non-perceptual sources, such as language, and for constructing models of stories and the like. Importantly, however, the child need not be aware of what is involved in this process. When it comes to evaluating a hypothetical model in terms of the reality model, the child need not explicitly realize that she is in fact engaged in an interpretive project involving the maintenance and comparison of multiple models.

It is the capacity to *represent* these interpretive processes, in addition to simply executing them, that constitutes level (3). Here's why. Let's put ourselves in the position of the child who has just seen Maxi return home after the chocolate has been moved. What the child at level (2) can do is to represent the real situation, which includes the fact that the chocolate has been moved. The child can also represent, at this stage, other possible models. She can think about what it might be like if the chocolate were moved upstairs, for example (and in this world, perhaps Maxi would go and look upstairs). But these abilities alone won't provide a solution to the false belief task. Just representing a non-actual situation isn't enough, obviously. And even representing such a situation and somehow 'associating' that representation with the agent (e.g. Maxi) won't work either. As Perner (who is discussing a false belief example involving a character, Mary, who is looking for a van and believes falsely that it is at the park) puts it:

The correct answer, of course, is that Mary will go to the park, and this answer can be found by matching her goal specification to her belief model. The decision to make that match raises, however, what I would like to call the *puzzle of false belief*. The puzzle is: If it is Mary's goal to achieve an outcome in the *real world*, why should a match for her goal specification be sought in the *counterfactual situation* described by her belief. After all, one would not do that in the case of pretense. If Mary *really* wanted to find the van, one would not assume she would look in a location where she merely *pretends* it is. (Perner, 1988)

The solution to this puzzle is for the child to realize explicitly that the agent uses this deviant model *to represent reality*. And this involves not only the ability to keep multiple models in mind, but in addition to model the fact that the agent will *interpret* her model as *referring* to the *real* world. In effect, it requires the ability to model the modeling process itself -- to have a model in which there are elements that are either models themselves, or employ and interpret models. We will return to this later.

5.3 Further thoughts on Leslie, Wellman, and the simulation theory

I will of course be arguing that Perner's account is correct, and can be seen as an application of ETM. Before I get into that (in the next section), however, I'd like to say some things about Leslie's metarepresentations, because I claimed in section 5.1 that Leslie's proposal was unsatisfactory and now is the time to discuss this. I have already talked a bit about pretend play in the previous chapter. There I argued that pretend play was the product of an altered control loop, similar to internal imagery in many respects.

Let's run through a typical example, such as pretending to fly an airplane. First we need some idea of what the real control loop would be when flying an airplane. Beginning with the controller, which I will take to be some CNS system, efferent commands will travel down the spinal cord to the arms and hands (at least). From there, the hands move various controls, such as a joystick, which in turn affects various devices on the plane, rudder, elevators, and the like. These in turn affect the environment surrounding the airplane, which in turn affects the aircraft -- its position, yaw, pitch, etc. These factors affect, through both instrumentation and unaided vision, the sensory apparatus of the pilot, who then uses this updated information while continuing to issue commands.

Now if I were to *imagine* flying a plane, one way would be to cut off the efferent control loop at the spinal cord (thus my arms and hands won't actually do anything), and resume it at some level of internal visual processing. I literally imagine, while remaining motionless with eyes closed, the experience of flying an aircraft. My internal emulator, which is not very good because I am not a pilot, thus takes over for the control loop between the efferent commands leaving via the cortico-spinal tract and the afferent signals coming in via VI. Everything is now internal. There is nothing sacred about these two points, however. I could keep my hands in the loop, and move them as if I were operating

a joystick, effectively pushing the real part of the efferent loop out a bit. I could even have something in my hands, like a banana, when I do this -- thus *pretending* that the banana is a joystick. A flight simulator allows one the option of breaking the control loop up at points external to the skull, as it emulates everything in between the controls and instrumentation. This allows the normal control loop to extend away from the body in *both* directions, which makes it that much easier to *pretend* that the giant machine is an airplane.

This seems to me to be the correct analysis of pretend play. By contrast I want to say that representing people as representers involves representing them as employing internal emulation. It is in a sense a recursive application of emulation. More on this later, but for now the important point is that this is qualitatively different from simply representing an altered control loop. In support of this, notice that pretend play typically involves the maintenance of some activity. It is quite a stretch for one to just put a banana on a table, carry on with some unrelated activities, and claim to be pretending that the banana is a telephone. Intuitively the act of pretense involves the active maintenance of a pattern of activities within which proxies can be incorporated. Leslie's account, however, makes no such prediction. Since the language of thought employs static representations which are merely 'decoupled,' the notion of an ongoing activity being crucial for such decoupling is hard to make sense of. Mentalese representations are purely logical in character, not tied to activity at all.

In addition to discussing control loops and their potential re-routing, I also had the opportunity in Chapter Four to argue that *theories* are in fact best understood as articulated emulators. In brief, the domain the theory is about will be the target domain, the articulators will be the theory's ontology, and the way in which the articulators interact will be the theory's laws. If this is true, and if folk psychology (and its ontogenetic predecessors) are in fact theories (as Wellman explicitly argues), then a change in the theory of human behavior amounts to a change in the way humans are emulated, and this in turn amounts to

a change in the articulants of the emulator and the processes of their interaction. I will be expanding on this presently.

Finally, I should make some mention of the relation between the position I am presenting here and the *simulation* theory of folk psychology. Insofar as I understand them, simulation theorists (such as Robert Gordon (1986), and Alvin Goldman (1993)) maintain that our understanding of the behavior of others comes in large part from our simulation of them, and specifically our ability to put ourselves in their place (in some sense, see Gordon (1986) for a discussion of what this might mean). I can predict that Maxi will look in the kitchen for the candy because I know that if *I* saw the candy hidden in the kitchen, and *I* didn't see it get moved anywhere, then *I'd* look for it in the kitchen.

Even though the terms 'emulation' and 'simulation' are otherwise quite similar, in this case I think they are not necessarily so. They may be compatible, but I will focus on what I take to be a potential difference. The simulation theory assumes that children can understand false belief in their own case, because it is only on that assumption that the child can be expected to gain any insight into Maxi's behavior by simulating Maxi's situation.⁵² There is thus a priority (certainly an explanatory one, and presumably an ontogenetic one as well) that understanding of the behavior of oneself has over understanding the behavior of others. On my account there is no such asymmetry in principle.⁵³ What one gains at a certain age is the capacity to represent the process of representation itself, and this capacity underwrites the explanation of behavior in others and oneself equally.⁵⁴

⁵²And there is some evidence to the effect that they don't. See Astington and Gopnik (1988).

⁵³In Chapter Seven, I will have reason to explain certain phenomena as resulting from the 'emulation' of another subject, and in these instances, I will mean something akin to simulation. However, the point at that stage will be content attribution, and not psychological explanation.

⁵⁴The advantage that first person explanations seem to have comes from more direct and complete knowledge of the contents of the model of reality, not from any qualitative difference in the understanding of how self and others *use* models.

5.4 Emulating emulators

I will attempt to translate Perner's program into ETM vocabulary. What Perner calls level (1), the simple *having* of a model, obtains where a system has an emulator that is used for on-line control purposes, such as providing faster feedback (see section 3.1), or enhancing sensory processing (as in section 2.4). The 'higher level' emulators will simply be unnoticed representations of the external situation, best thought of as being modulated by the senses, as explained in section 3.4. What Perner is calling level (2), the *use* of a model, will correspond to a situation where a system has an emulator that can be decoupled for purposes of imagery (section 3.3) and for considering hypothetical situations (2.4). This is represented in Figure 5.1. In this diagram, the brain has a controller (marked '*'), which interacts with the world via efferent pathway A. The world provides feedback via C. The internal model is driven by efferent copy information from B, and is kept updated by feedback from the senses, via C-E. The controller's activity is governed by feedback path F, which is a combination of emulator output D, and 'real world' feedback C.

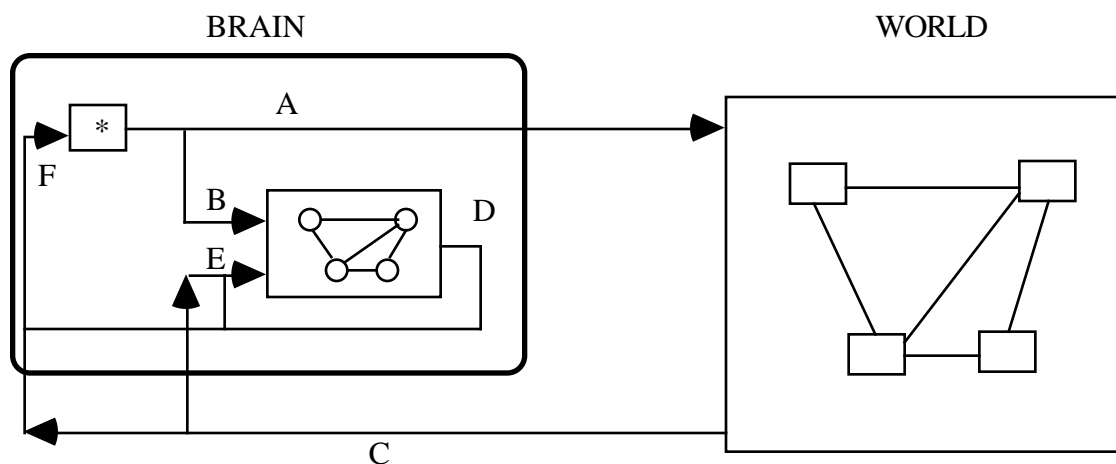


Figure 5.1: Schematic of brain emulating the external world.

In diagram 5.2, however, we have a more sophisticated situation. Here the first cognizer, call her **A**gent, is contemplating the activities of a second cognizer, **B**gent. In the objective situation Bgent is interacting with an environment and is representing it via a constantly updated emulator. Agent, on the other hand, is observing Bgent. Agent will also be keeping track of what is going on by updating a world model, but this world model is more sophisticated, in that it includes, as separate articulants, representations of Bgent as well as of Bgent's internal model, and of the semantic relationship between the two, that Bgent interprets her model *as* reality.

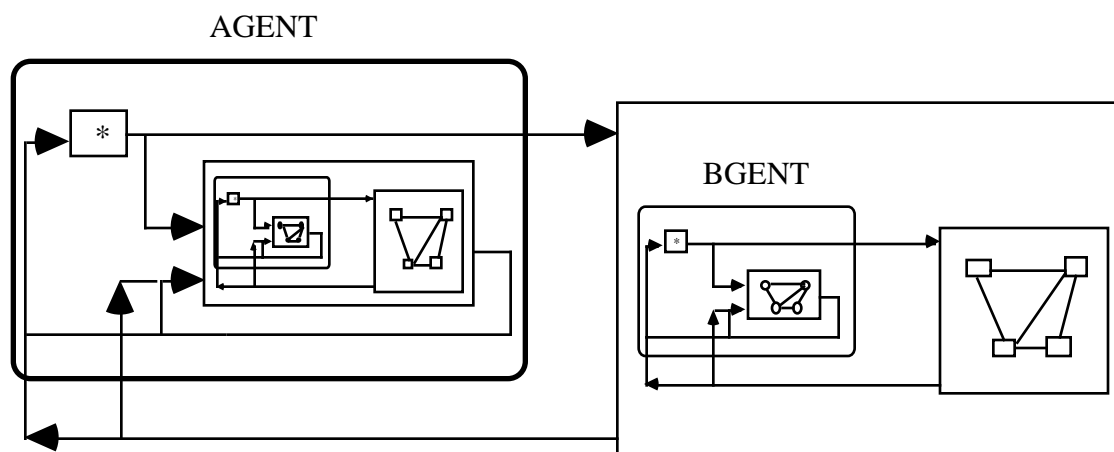


Figure 5.2: Schematic of brain emulating an agent *as* emulating the world.

In Chapter Seven, I will explore in greater detail the semantic issues raised here, and will show that this account of the theory of mind will shed significant light on a host of troubling phenomena associated with semantics, such as opacity.

5.4 Conclusion

In effect I have simply commandeered Perner's model wholesale. I am almost too embarrassed about this chapter to include it. This project has enough of the flavor of a

long-winded annotated bibliography as it is, and I haven't done anything here to dispel that interpretation. But there is a more optimistic way to look at the payoff of this chapter, which is as a demonstration of the rather seamless and natural way in which the strategy of emulation, initially motivated in the domains of *motor control* and *imagery*, can be easily extended to understanding the prima facie quite distinct domain of the *development of the theory of mind*. I think when viewed in this more holistic way, the lack of tortured argumentation and absence of the need for original bells and whistles can be seen as the virtue it truly is. Those who like tortured argumentation and lots of original bells and whistles are invited to proceed to Chapter Six.

Chapter Six: The Grammar of Thought

Semantics casts a syntactic shadow.

John Searle

The story as I have developed it so far maintains that at least in some cases brains construct and configure emulators for a range of purposes, including planning, considering counterfactuals and the support of imagery. We emulate, for purposes of prediction and understanding, a variety of domains, including the external world and its inhabitants. Furthermore, these emulators will in many cases be articulated to varying degrees, and will thus bear some resemblance to *models*. And finally, some of those entities which are emulated are themselves represented as capable of emulation. I will in this chapter argue that another use of emulators is to support communication. To put it crudely, one way for me to communicate to you information about how the world is or might be, is for me to reconfigure (aspects of) your reality emulator, or at least a temporary emulator that you allow to be configured by me for communicative purposes.

This idea, or versions of it, is not new. Researchers as otherwise diverse as Johnson-Laird (1983), Jackendoff (1990), Fauconnier (1985) and Langacker (1987, 1990, 1991) have embraced variants of this view, as have many others. One feature that these views, and my own, have in common is the claim that the semantics of natural language is parasitic on the semantics of kinds of mental or cognitive constructions. One difference between the present treatment and those just mentioned is that I am identifying the

meaningful mental construction that is the target of language as an emulator (as opposed to a model, or mental space -- though I will want to say that these are special kinds of, or idealizations of, emulators).

Furthermore, I will argue that the ways that these semantic cognitive constructions are maintained and manipulated will have grammatical consequences. This view I share with at least Langacker and Johnson-Laird. But since the semantic cognitive constructs in my account will be articulated emulators, and because I have already said some things about how articulated emulators are constructed and used, I will be able to make some specific predictions that will be borne out in the grammar of English. In short, the claim, which is as strong as it is inescapable, is that a proper understanding of language use requires not only a knowledge of the extensions of the words, a savvy about 'syntactic' phenomena, sensitivity to discourse constraints, and considerable reliance on encyclopedic world knowledge, but also on the way that nonce semantic constructions are configured, maintained and manipulated by cognitive processes.⁵⁵

6.1 Innateness and syntactic autonomy

One of the strongest and most influential hypotheses about the nature of human cognition is that it supports language use by means of innate, modular, language-specific, species-specific mechanisms. This doctrine is of a piece with the claim that syntax is by-and-large independent, at least at some level of processing, of both semantics and domain general cognitive constraints. It is this alleged selective sensitivity to the syntactic structure of language that is taken to provide the strongest evidence for an innate linguistic endowment. Consider the following:

⁵⁵I owe this particularly lucid way of expressing the thesis to Nili Mandelblit (1993, private communication), who has helped me to see more clearly how the issues of this chapter play a role in a rich context of linguistics and machine translation.

- (1) Colorless green ideas sleep furiously.
- (2) *Furiously sleep ideas green colorless.⁵⁶
- (3) Police police police police police.⁵⁷
- (4) The horse raced past the barn fell.
- (5) What what what he bought cost would buy in Germany would surprise you.

The sentence (1) seems to be acceptable in a sense in which (2) is not. The sense in question is a *syntactic* sense. Though (1) is semantically odd, it is well formed, it can be parsed and understood (for example, one could ask "What are the ideas doing?" and the answer is clearly "sleeping"). We are thus led to conclude that syntactic properties of sentences are independent of their semantic properties.

Examples such as (3) and (5) are meant to demonstrate the independence of syntactic competence from syntactic performance, and thus to demonstrate that syntax is independent of general cognitive constraints, such as attention and short-term memory. Thus when given the appropriate hints, (3) is seen to be grammatical, and if given enough time, and maybe a pencil and paper, most people could determine that (5) is in fact grammatical, though certainly atypical. This initial impression of senselessness is a performance limitation on processing multiply embedded clauses. The point is that (3), like (4), is grammatical. The fact that they may not seem so is purely a performance problem, and does not render them ungrammatical.

Once the premise that syntax is independent of semantics and cognitive constraints is allowed, though, the cause has been lost. It turns out that when we in fact try to explicate

⁵⁶ I will follow these conventions established within the linguistics literature: Sentences that are cited to show the grammaticality or ungrammaticality of some construction will be unmarked (grammatical), preceded by an asterisk '*' (ungrammatical), or preceded by a question mark '?' (questionable).

⁵⁷ I owe this example to David Kirsh. One way to parse it is as syntactically identical to 'Cats dogs chase pursue mice.'

the syntactic structures of a language without invoking semantics or general cognition, we are forced into positing exceedingly complex and sophisticated processing, of a sort that could not conceivably be learned on the basis of exposure to language alone. Constraints on which of these syntactic structures can be learned must be posited, and these constraints must be purely syntactic in nature (semantic or cognitive constraints won't do the trick because syntax is independent of these domains). We are forced to conclude that humans have an innate linguistic endowment that is language specific (indeed, syntax specific), and which operates by, at a minimum, constraining the space of possible syntactic theories enough to allow exposure to language to be sufficient for learning.

The previous argument has been very influential, most notably in the work of Chomsky, Fodor, and their followers. Nonetheless, I intend to show that this argument, and others like it, are in fact not so strong as they seem. In particular, I will argue that semantics and domain-general cognitive processes (provided these are correctly understood) are sufficient to account for data that have been assumed to require powerful autonomous syntax. Obviously a complete account of all such phenomena is beyond the scope of this chapter, so I will content myself with two more modest goals. First, I will provide analyses of a few representative linguistic phenomena. Second, I want to provide reason for thinking that ETM is compatible with Ronald Langacker's Cognitive Grammar, which itself rejects the formalist/domain-specificist posits of formal syntax, and which has been used to address a fairly wide range of 'syntactic' data.

ETM maintains that cognition makes use of emulation, of the internal simulation of some target domain or other. In particular, these simulations can be of aspects of the external world, or more correctly of the external world as typically interacted with. The articulators in such cases will mirror recurring relations, activities and objects which play a useful organizational role for the agent in her dealings with the environment. If natural language is seen as sets of instructions for the configuration of emulators, then constraints

on what emulators can and cannot do ought to show up as constraints on what sentences are and are not grammatical. In turn, these constraints can be of two forms. Cognitive constraints arising from how emulators are neurally implemented, and semantic constraints⁵⁸ arising from how emulator articulators are most felicitously or easily orchestrated.

The next section of this chapter will explore some linguistic phenomena that center around anaphoric relations and extractions (these will be explained in due course). I will first provide sketches of the currently received account within Chomskian formal linguistics. This will be in the Government and Binding framework (henceforth GB, see Chomsky 1981, 1986). Next, I will demonstrate some persistent problem cases for the GB account. Most interesting will be pairs of sentences which are *syntactically* identical, and yet have different *acceptability* status, seemingly as a function of semantic or cognitive features. Having raised difficulties for the GB treatment of anaphora and extraction, I will advance a framework (ETM as developed in previous chapters, with a number of crucial refinements and additions) that addresses the phenomena in a much simpler and more illuminating way.⁵⁹

6.2 Anaphora and extraction

Anaphora and extraction are phenomena in which, *inter alia*, one element in a sentence is linked to some other element in the sentence, usually in the sense that the linked elements have the same referent. The most obvious cases are pronouns and reflexives,

⁵⁸ This initial distinction is compatible with the possibility, to be explored in later chapters, that the distinction is superficial.

⁵⁹ The treatment developed owes much to Ronald Langacker's Cognitive Grammar framework (henceforth CG, Langacker 1987, 1991). So much, in fact, that I will commonly adopt his terminology and notation in this chapter.

where a pronoun is taken to have the same referent as some other noun phrase in the sentence, such as

(6) Jerry_i went to the party, and he_i didn't leave until the next morning.⁶⁰

In (6) the pronoun 'he' and the NP 'Jerry' are taken to corefer. However, consider the following:

- (7) *Jerry_i understands him_i.
- (8) Jerry_i understands him_k.
- (9) Jerry_i understands himself_i.
- (10) Jerry_i believes that he_i is right.
- (11) *Jerry_i believes that himself_i is right.
- (12) Jerry_i believes himself_i to be right.
- (13) *Himself_i/*He_i believes Jerry_i to be right.
- (14) Before he_i went to the store, Jerry_i phoned the office.

The coreference renders (7) ungrammatical, as can be seen by the grammaticality of (8). (7) can be fixed, however, by replacing the pronoun 'him' with the reflexive 'himself,' as in (9). In (10) and (11), the opposite relationship holds -- use of the pronoun renders (10) grammatical while use of the reflexive renders the otherwise identical (11) ungrammatical. In fact, pronouns and reflexives have mutually exclusive distributions.⁶¹ That is, if a pronoun, coreferential with some NP, is grammatical in some location,

⁶⁰ In the remainder of this chapter, subscripts will be used to indicate coreference (same subscript) or non-coreference (different subscript).

⁶¹ This is not completely true. There are some notable exceptions, such as 'picture NPs', which, though fascinating and potentially quite illuminating, are beyond the scope of this chapter.

substituting a reflexive with the same referent in the same location will be ungrammatical, and vice versa. A look at (10), (11) and (12) will indicate that it is not immediately obvious how these distributions are determined. For example, on the basis of (12) and (13) one might assume that the anaphor must *follow* the coreferential NP, but (14) shows that this cannot be right.

Another sort of phenomenon, which, on the GB account, also has to do with coreference, is extraction. This occurs when some constituent is moved from its normal location to some new location in the sentence, normally during the transition from d-structure to s-structure.⁶² Consider:

(15) Jerry saw a cow.

(16) Jerry saw what?

(17) What did Jerry see?

Ignoring the antics of the auxiliary 'did', what happens in (17) is that a questioned constituent, manifested at s-structure as 'what', is moved from its sentence final position in (16) to a sentence initial position in (17). 'What' is known as a 'wh-word', and the movement described is a variety of wh-movement, or wh-extraction. To a first approximation, extraction involves the motion of an extracted phrase from an extraction site to a target location, and in order for the sentence to be processed correctly, one must be able to reconstruct the extraction site and correlate the features of that site with the extracted phrase. For example in (17) the referent of 'what' must be understood to be the direct object of 'see', as opposed to an anaphor for Jerry, or whatever. GB (like many other flavors of formal syntax) addresses this issue by claiming that the extraction site still

⁶² D-structure and s-structure are posits of Government and Binding theory, and are analogous, but not identical, to deep-structure and surface-structure in transformational accounts.

contains a non-overt element, a 'trace', which is linked to the wh-word. Thus the sentence could better be represented as

(18) What_i did Jerry see t_i?

The trace is of course non-overt, i.e. phonetically null. Co-indexation brings out one analogy with anaphora, another being that the acceptability of extraction is seemingly subject to non-obvious syntactic constraints.

(19) The woman with the hat ate salmon.

(20) *What_i did the woman with t_i eat salmon?

(21) What_i did the woman with the hat eat t_i?

Apparently, as (20) shows, the extraction from the adjunct PP 'with the hat' is not allowed. The question now is: How do we provide an account of the patterns of acceptability and unacceptability manifested here? We will briefly tour the GB account in the next section. This formal approach promises to provide an explanation of the difference between e.g. (20) and (21), and thus to provide insight into the content of the innate language-specific mechanisms we all share.

The sentences in (22) - (24) illustrate the final phenomenon that will interest us.

(22a) Jerry met the referee yesterday.

(22b) *Jerry met yesterday the referee.

(22c) Jerry met yesterday the head referee from Superbowl XXI.

(23a) John introduced Mary to Linda.

(23b) *John introduced to Linda Mary.

(23c) John introduced to Linda the president of the company that just published her book.

(23d) ?John introduced the president of the company that just published her book to Linda.

(24a) I drink water every day.

(24b) *I drink every day water.

(24c) I drink every day two rum and cokes with a twist of lemon.

The surprising patterns are as follows. The (a) sentences illustrate the prototypical occurrence of verb with two adjuncts - a complement and a modifier ((22a) and (24a)), and 'introduce' with two complements (23a). When the order of the adjuncts is switched, felicity is lost (the (b) sentences). However, when the NP adjunct which is normally closest to the matrix verb is sufficiently 'heavy' (long or complex), it can legitimately be switched with the other adjunct. In fact, as (23d) shows, sometimes this shift is somewhat mandatory to avoid marginal sentences.⁶³

Heavy-NP Shift, as this phenomenon is called, receives considerably less attention in the GB literature than the previous two phenomena, as I quickly discovered when trying to do a literature search on the topic. Perhaps this is because it is something of an embarrassment. At all levels of syntactic representation (d-structure, s-structure, LF) the relevant features of the (b) and (c) sentences are identical (or so it would seem, see section 6.3), and yet native English speaker acceptability judgments differ robustly. Crucially, it seems difficult to explain it away as a performance problem. A performance problem should be expected to manifest itself by making certain sentences which are *formally* acceptable seem *ungrammatical*. It is quite different when one has a normally

⁶³Surprisingly, some people to whom I've shown this data claim not to buy it. I suspect that this feeling results from explicit reflection on the construction, and would not hold up in unreflective processing. If you read past the sixth sentence in this chapter (where the NP following 'to you' has been shifted), consider yourself a competent HNPS processor.

ungrammatical construction become *acceptable* when the processing load is increased. Before going into the positive account of these linguistic phenomena, it will be instructive, at least for purposes of contrast, to take a brief look at how they are handled by current formal syntactic theory.

6.3 Outline of the GB account

In this section I will outline a GB account of the three phenomena discussed above. My goal is not to provide a comprehensive analysis of these phenomena within the GB tradition. Nor will I, later in the chapter, provide detailed argument against all aspects of the GB account here outlined. Rather, I simply want to provide some idea of how current formal syntactic theory approaches the problems at hand.

Let us look more closely at (7) and (8), repeated here as (25) and (26):

(25) *Jerry_i understands him_i.

(26) Jerry_i understands him_k.

(27) Jerry_i understands himself_i.

(28) *Jerry_i thinks everyone understands himself_i.

A tree diagram⁶⁴ of (27) is provided in Figure 6.1, and (28) is diagrammed in Figure 6.2.

⁶⁴ In all such diagrams I will assume, as is commonplace in GB, X' Theory (pronounced 'X-bar'), which features only binary branching. For more on X' Theory, see Haegeman (1991), Stuurman (1985).

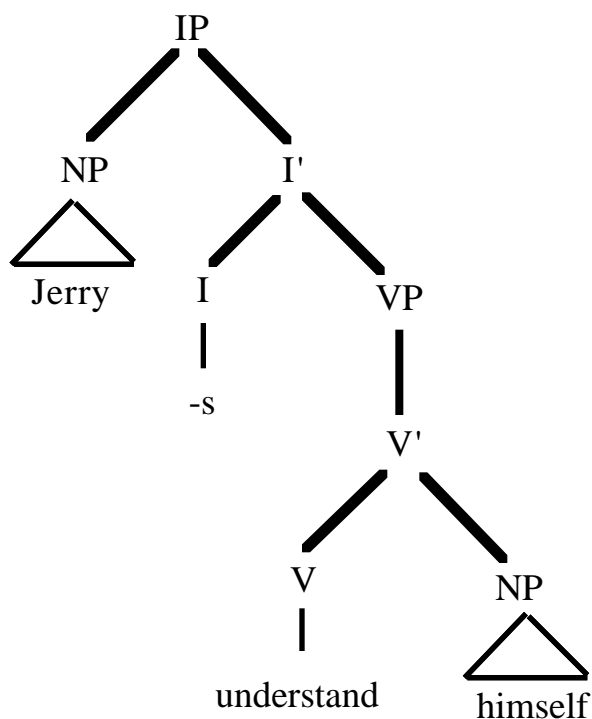


Figure 6.1: 'Jerry understands himself.'

The important feature of this diagram is that we can easily describe a relation, called c-command, which plays a part in determining the distribution of anaphoric pronouns⁶⁵ and reflexives. To a first approximation, A c-commands B iff the first branching node dominating A also dominates B, and neither A nor B dominates the other. In (27) clearly the subject NP c-commands the object NP 'himself'. The first branching node dominating 'Jerry' is IP, and IP does dominate 'himself'. The reverse does not hold. The first branching node dominating 'himself' is V', which does not dominate 'Jerry'.

⁶⁵ An anaphoric pronoun is one that has an antecedent in the sentence. Not all pronouns corefer with an NP present in the same sentence.

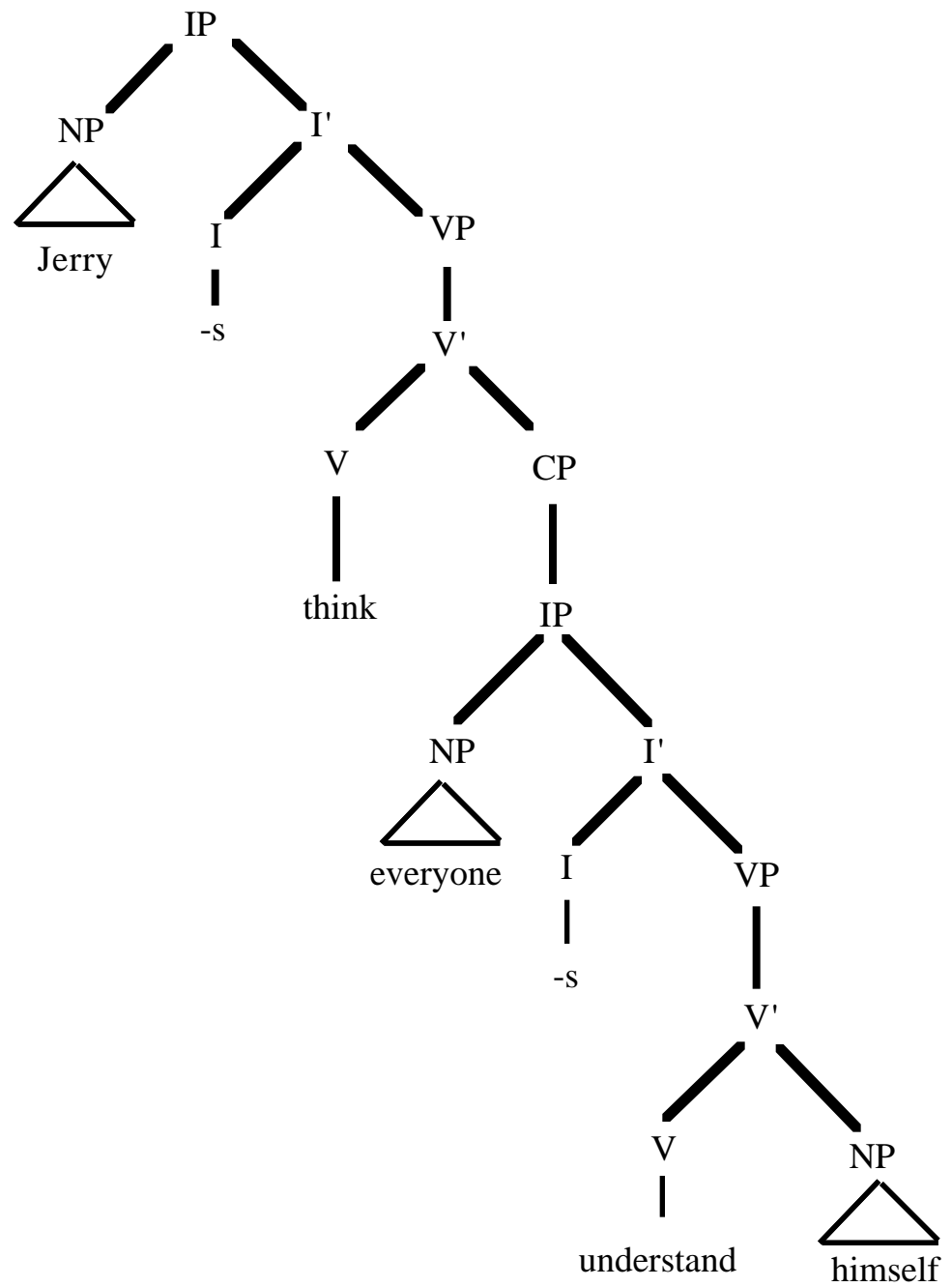


Figure 6.2: 'Jerry thinks everyone understands himself.'

We can now state the requirements on the distribution on reflexives and pronouns:

(29) Reflexives (and reciprocals) must be bound in their governing category.

(30) Pronouns must be free in their governing category,

where A binds B iff A c-commands B and A and B are co-indexed.

A governing category for an anaphor or pronoun is the minimal domain containing it, its governor(s), and a subject.

A governor, to a first approximation, is the head that case marks the governee. In this case, the verb 'understand' governs the reflexive and assigns it accusative case. For present purposes, a subject can be taken to be the nominal subject of a finite clause.⁶⁶ Thus in (25), the governing category must include the entire sentence. The governor in this case, (27) is 'understand' which assigns accusative case to 'himself'. The subject is 'Jerry', and since the reflexive is bound (i.e. it is coindexed with and c-commanded by 'Jerry'), everything is OK.

But now let's look at (28). Here the governing category is limited to the finite subordinate clause 'everybody understands himself'. This clause contains the reflexive, its governor ('understand') and a subject ('everybody'), and thus the reflexive must be bound within this clause. But it is not. It is bound by 'Jerry', which still c-commands the reflexive and is co-indexed with it, but 'Jerry' is not within the governing category. Hence the ungrammaticality. If a pronoun replaces the reflexive, however, then it will be free (i.e. unbound) in its governing category, and thus the sentence will be grammatical. Likewise, if the reflexive is coindexed with 'everyone', then it is bound in its governing category, and all is again well.

⁶⁶ Though a complete account must recognize a substantially broader notion of what is here called 'subject', to include verbal agreement features, and even 'subjects' of NPs, such as 'John' in 'John's invitation to the Smiths.' See Haegeman (1991) for more detail.

We have here a glimpse of the formalist machinery in action. The distribution of reflexives and pronouns is explained in a manner that makes no reference to semantics, or to general cognition. We appeal only to structural constraints, such as c-command, and purely syntactic entities, such as governors and subjects.

Extraction

The constraints on *wh*-movement derive mainly from the relationship between the target location (i.e. the moved phrase in its new sentence initial location) and the trace left at the extraction site (as well as any intermediate traces). In effect, certain syntactic elements, like maximal projections (e.g. full noun or verb phrases), can impede the movement of an element from its extraction site to the target location. This much has remained fairly constant throughout the last 15 or 20 years of syntactic theorizing. What has changed is that in previous paradigms, the definition of an 'impeding element' was fairly clear, but certain additional constraints had to be added to account for various exceptions (such as a 'that-trace' filter being added to a relatively straight-forward notion of subjacency). On the other hand, in the more modern version expounded in Chomsky (1986), the impeding elements, or 'barriers' (hence the title of the work), are given fairly complicated definitions, but with the putative advantage that with the more refined notion of a barrier, there need be no appeal to additional filters or rules.

In this section I will give a very brief overview of Chomsky's account (as developed in Chomsky (1986)).⁶⁷ Movement of an extracted phrase leaves a trace, and in many cases, movement over long distances will take place in cycles, where during each cycle the moved phrase leaves another trace. When the movement is complete, there will then be a chain, with the initial trace at the extraction site at the foot of the chain, and the

⁶⁷ This exposition will be greatly simplified, as even a reasonable first approximation to the received account is quite beyond the scope of this section. I would refer the interested reader to Chomsky (1986), Haegeman (1991, chapter 10) and Rizzi (1990) for more complete treatments.

moved phrase itself at the head, and in between a series of intermediate traces, if any. Constraints on *wh*-movement manifest as constraints on where moved phrases may land during their trek across the sentence, and as structural constraints that must obtain between any two adjacent traces.

Movement must satisfy the subjacency condition, which is that any two adjacent elements in the movement chain must be *subjacent*, and two elements are subjacent iff movement from the first to the second (i.e. in the direction from the foot to the head of the chain) does not cross more than one barrier. Deceptively simple. Now to what counts as a barrier. First, we must consider a construct called a 'blocking category', or BC.

(31) P is a BC for Q iff

- i) P is a maximal projection (NP, IP, CP, etc.) that dominates Q, and
- ii) P is not L-marked.

The idea is that the element Q is being moved over P, and we want to see if P is going to be a barrier for this putative movement. Condition (31i) is straightforward. As for (31ii), an element is L-marked iff there is a lexical category (i.e. N, NP, PP, but not C or CP) which theta-governs it.⁶⁸ Now that we have determined what the BCs along the movement path are, we are in a position to find the barriers.

(32) X is a barrier for Y iff either (i) or (ii)

- (i) X is a maximal projection which immediately dominates a BC for Y.
- (ii) X is a BC for Y, and X is *not* IP.

So, in order to find the barriers, we first determine all the maximal projections along the movement path, and eliminate those that are L-marked (by (31i)). We then split

⁶⁸ An element A theta-governs another element B iff (roughly) A governs B, and A assigns a theta-role to B. The criteria for government are as complicated as those for subjacency, but to a first approximation can be defined as mutual c-command. As for theta-roles, certain elements require other elements as part of their meaning. For example, the verb 'hit' requires an agent and a patient, and we can say that this verb assigns these theta-roles (for 'thematic role') to these elements. The verb 'believe', for example, assigns an agent role, the believer, as well as a theta-role to a complementized sentence (CP), such as 'that the cat is on the mat'. Thus, the CP in 'John believes that the cat is on the mat' is not a BC, since it is theta-governed by the verb 'believe'. Thus movement out of CP, such as 'What does John believe the cat is on t?' crosses no barriers.

the remaining maximal projections into two groups, IPs and non-IPs. The non-IPs are barriers (by (32ii)) inherently, and the IPs are barriers if they immediately dominate a BC (by (32i)). Thus IPs can be barriers by 'inheritance'.

And finally, we can determine if a given *wh*-movement is legal. Starting at the extraction site, we follow the movement to the next landing site (the next intermediate trace). If this movement crosses no barriers, or one barrier, then it is legal (by the subjacency condition). We continue this way until the entire chain has been checked. If all the movement cycles are legal, then the sentence is grammatical (at least as far as the movement in question is concerned).

(33) The woman with the hat ate salmon.

(34) *What_i did the woman with t_i eat salmon?

(35) What_i did the woman with the hat eat t_i?

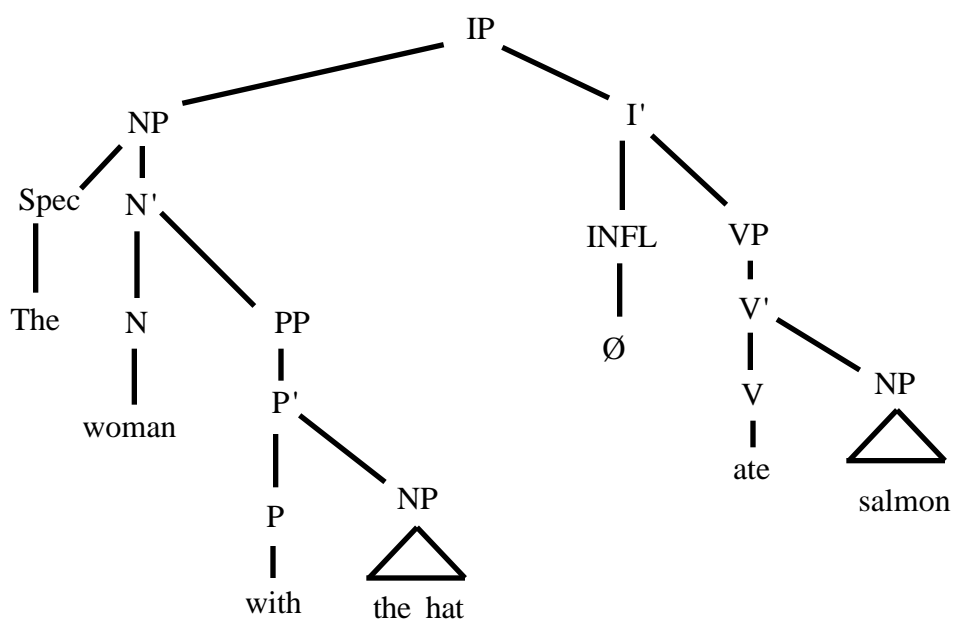


Figure 6.3: 'The woman with the hat ate salmon.'

Consider (19) and (20), repeated here as (33) and (34) and diagrammed in Figure 6.3 and Figure 6.4. Notice that the NP 'the hat' cannot be moved out of the NP 'the woman with the hat', since doing so would cross two barriers (PP and NP). In fact, the movement to the [Spec, CP] position would require movement across four barriers: PP, NP, IP (by inheritance for immediately dominating NP), and CP.

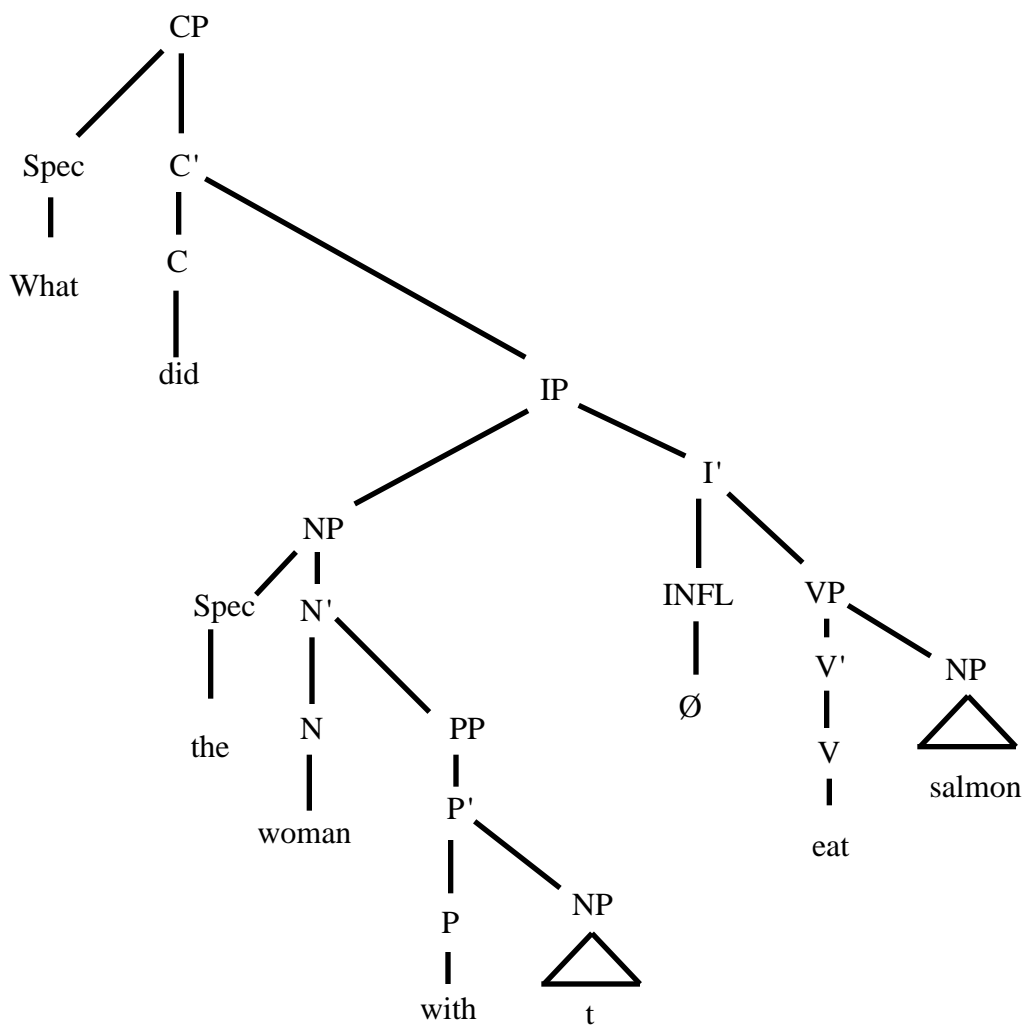


Figure 6.4: 'What did the woman with eat salmon?'

One might initially assume that movement of the NP 'salmon' to [Spec, CP] would also cross too many barriers, two in this case, VP and CP (note that IP is not a barrier in this instance since it does not immediately dominate any BC). Grammaticality is 'saved', however, by positing that the movement actually takes place in cycles, first from the extraction site to a site adjoined to VP, and from there to [Spec, CP], as in Figure 6.5.

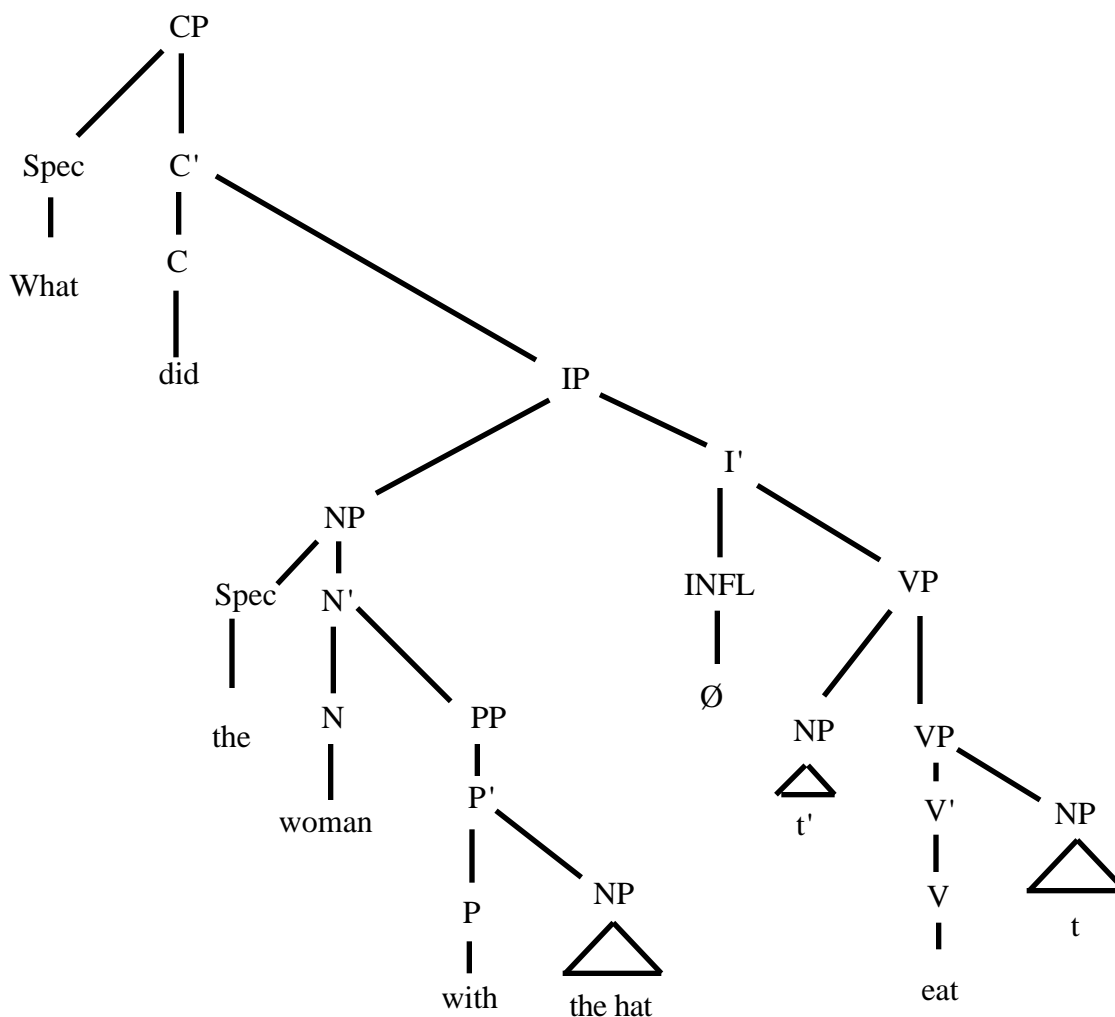


Figure 6.5: 'What did the woman with the hat eat?'

The (arguably dubious and ad hoc) 'explanation' of the grammaticality of (35) is that the first movement cycle moves the NP from the extraction site, and adjoins it to VP, to make another VP.⁶⁹ The second cycle involves movement from this adjoined position to the final site at [Spec, CP]. By stipulation, movement out of the VP adjunction does not count as crossing a barrier, hence the second movement cycle crosses only one barrier, CP, and is thus legal.

Heavy NP-Shift

As with anything else, there can hardly be said to be agreement within the GB community as to the correct treatment of Heavy NP-shift. I here present briefly the

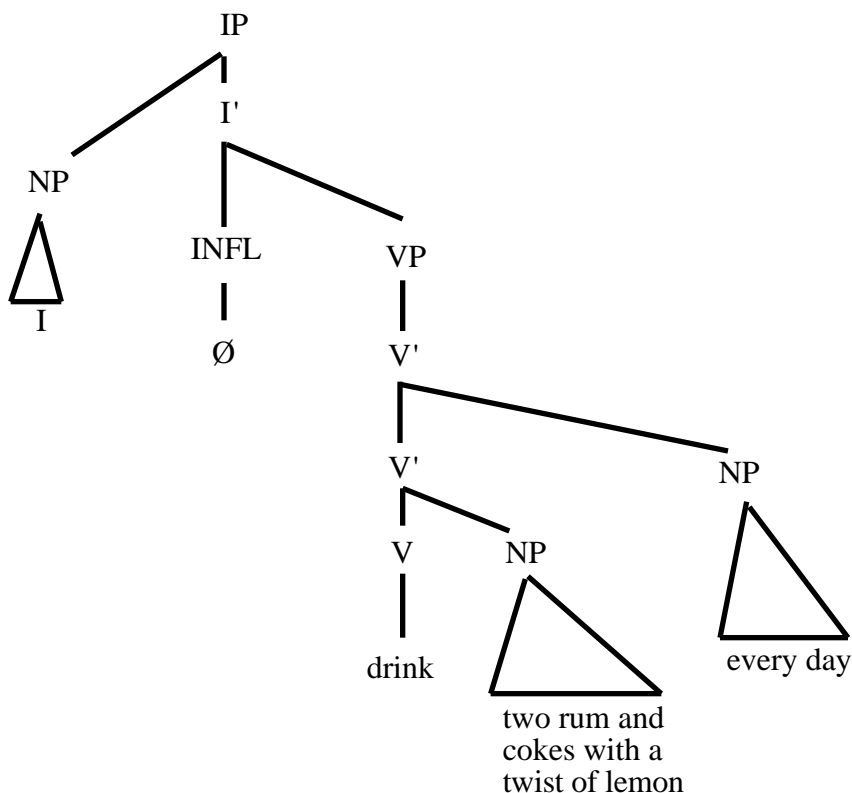


Figure 6.6: 'I drink two rum and cokes with a twist of lemon every day.'

⁶⁹ VP adjunction is putatively justified in Chomsky (1986) for reasons involving quantification at LF. Haegeman (1991) also points out the VP adjunction is 'independently motivated' by Heavy NP-Shift. This last claim will be explored in the next subsection.

treatment offered in Haegeman (1991). In short, the story is that at d-structure, the constituents are generated in their normal positions, but that in the transition to s-structure, the heavy constituent is moved to a new VP-adjoined position "created for it." Thus the d-structure of (24c), repeated here as (36) is (37), and its d-structure is diagrammed in Figure 6.6, and s-structure in Figure 6.7.

(36) I drink every day two rum and cokes with a twist of lemon.

(37) I drink two rum and cokes with a twist of lemon every day.

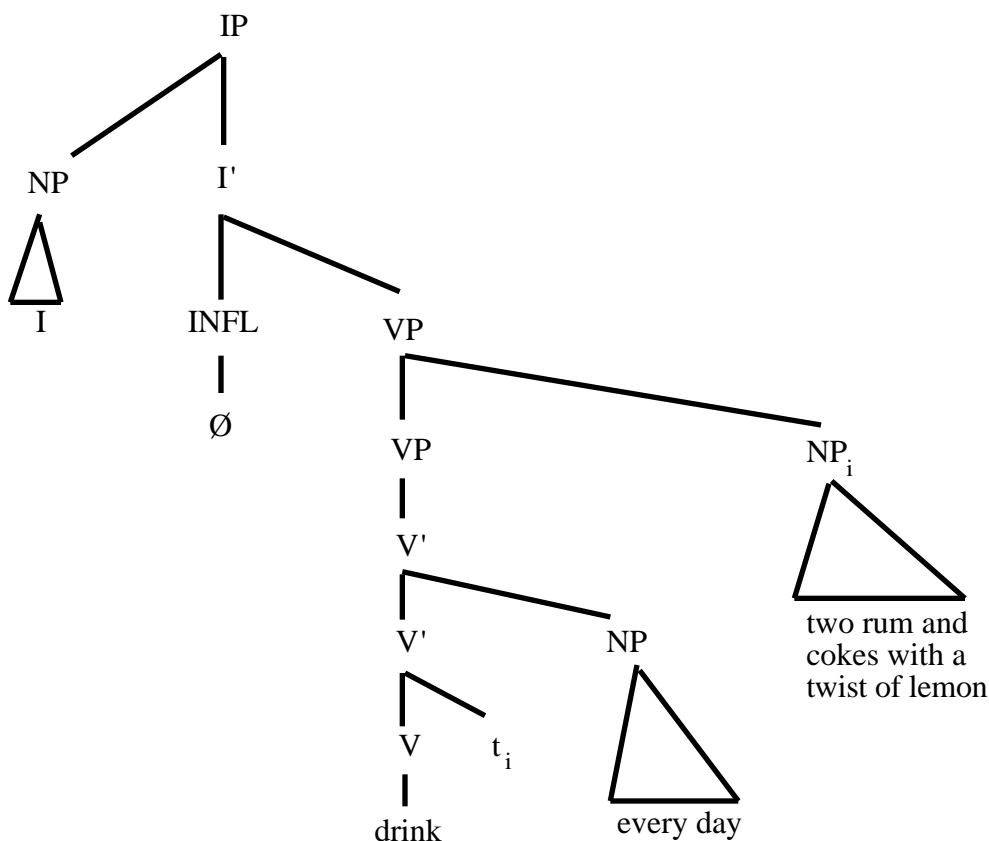


Figure 6.7: 'I drink every day two rum and cokes with a twist of lemon.'

Problems for GB and autonomous syntax

Appearances notwithstanding, all is not well with formal syntax. One sign of trouble is the reluctant admission of a distinction between 'core grammar' and constructions on the 'marked periphery.' This distinction is forced by a recognition that the principles of so-called Universal Grammar (UG) cannot account for all the acceptable constructions in a language.

(38) I liked the gears in *that* car.

(39) Which car_i did you like the gears in t_i?

(40) I liked the car with the *new* gears.

(41) *Which gears_i did you like the car with t_i?

(42) Which NPs_i are there unusual possibilities for extraction from t_i?⁷⁰

(43) Noam found, and Jerry described, a language module in the brain.

(44) What_t did Noam find t_i and Jerry describe t_i?

(45) *What_t did Noam find a language module in the brain and Jerry describe t_i?

(46) Karen bought a car_i and gave it_i to Susan.

(47) *What_t did Karen buy a car and give t_i to Susan?

(48) Karen went to the store and bought a muffin for Susan.

(49) What_t did Karen go to the store and buy t_i for Susan?

(39) shows extraction of an NP ('that car') from within a matrix object NP ('the gears in that car'). Notice, though, that the syntactically identical extraction in (41) is ungrammatical. The reader is invited to check that the extraction in (42) violates the

⁷⁰This is the title of Deane (1988).

constraints outlined earlier in this chapter. In (43) - (47) we see the across-the-board constraint on extraction, which seems to require that elements can only be extracted from a coordinate structure if they are extracted from all conjuncts. Again, as (49) shows, this constraint is not always obeyed. More examples of this nature will be introduced shortly, but for now the point is that autonomous syntax is not without problems, since we have syntactically identical pairs of sentences which nonetheless differ sharply in their acceptability, and these are not good candidates for performance errors (such as 'Police police police police').

The retreat position is that UG allows for the learnability of a core grammar, but in addition each individual's linguistic productions will include constructions that do not conform to this core.

But it is hardly to be expected that what are called "languages" or "dialects" or even "ideolects" will conform precisely or perhaps even very closely to the systems determined by fixing the parameters of UG. This could only happen under idealized conditions that are never realized in fact in the real world of heterogeneous speech communities. Furthermore, each actual "language"⁷¹ will incorporate a periphery of borrowings, historical residues, inventions and so on, which we can hardly expect to - and indeed would not want to - incorporate within a principled theory of UG. For such reasons as these, it is reasonable to suppose that UG determines a set of core grammars and that what is actually represented in the mind of an individual even under idealization to a homogeneous speech community would be a core grammar with a periphery of marked elements and constructions.
(Chomsky 1981, pp. 7-8)

In addition to this theory-internal distinction, formal syntax can also make use of fudge-room supplied by theory-external choices concerning what data are taken to be important:

⁷¹ It is interesting that twice in this passage Chomsky feels compelled to put scare-quotes around the word 'language' when used to indicate an actual spoken human language, as if in his ideolect, the word 'language' does not, strictly speaking, apply to such things.

With regard to the theory of movement, it appears that a number of different factors enter into informant judgments, including lexical choices to which they are sensitive; it is therefore necessary to proceed on the basis of some speculations concerning the proper idealization of complex phenomena, and how they should be sorted out into a variety of interacting systems (some of which remain quite obscure), which ones may tentatively be put aside to be explained by independent (sometimes unknown) factors, and which others are relevant to the subsystems under investigation.

(Chomsky 1986, p. 1)

Admittedly, these cautions and hedges are typical of scientific investigation, and so we should not condemn Chomsky's tradition simply on their account. However, the frequent appeal to such loopholes opens the possibility that formal syntax is just a false theory that happens to enjoy a measure of formal overlap with a strikingly different and much more complete account. An appropriate analogy would be between the abilities of general relativity and the crystalline spheres to provide observationally adequate predictions concerning the motions of the planets. This analogy is perhaps unfair to the theory of the spheres, which, unlike GB, did a pretty good job while not ignoring any of the available data.

6.4 Cognitive grammar

In light of the *prima facie* shortcomings of the purely formal analysis, I suggest we begin constructing a different account of linguistic competence. This account will consist of a theory of grammar, Ronald Langacker's Cognitive Grammar (to be introduced in this section), which I will supplement in three ways. The supplements will be a) the identification of (at least some) emulator articulators with Langacker's Conceptual

Predicates, b) a theory of variable binding,⁷² and finally c) accommodation and exploitation of the temporal character of language processing. This account will be quite schematic. If, by the end of this chapter, I have convinced the reader that this approach holds any promise (and in fact, even if I haven't), I would recommend Langacker's corpus as the best source for more detail⁷³ (keeping in mind the supplemental components (a) - (c) mentioned above and described in subsequent sections).

According to Cognitive Grammar (henceforth CG), linguistic competence resides in a mastery of three types of construct: phonetic units, semantic units, and symbolic units (which consist of associated phonetic/semantic pairs). Phonetic units are learned patterns for the construction of phonological material. This includes specific material such as individual morphemes (e.g. plural '-s' in English) and lexemes ('dog'), as well as more complex units, perhaps including material that is specified only schematically (e.g. 'both ... and ...'). Semantic units are elements of cognitive processing. These will include conceptions (including conventionalized, perhaps schematic, images) of things such as cats and mats, of relations such as 'being on' and 'before', complex temporal relations such as 'into' and 'across', and hosts of more subtle elements of cognitive processing, such as change of perspective ('The bank rises steeply from the river' vs. 'The bank drops steeply to the river'), alterations in conceived time ('go' vs. 'gone'), comparisons to implicit norms ('I got a little sleep before the exam' vs. 'I got little sleep before the exam'), etc.

Like a mental models account, CG sees the operation of language as the communication of meaningful cognitive structures that are not propositional in character, but rather model-like or image like (though perhaps including other sorts of structures as

⁷² One of the dangers of multidisciplinary work is terminology clash. Binding is an aspect of Government and Binding theory (appropriately enough), the aspect which handles coreference properties among elements. The other sort of binding, to be discussed later in this chapter, has to do with how features of objects are cognitively bound together as aspects of a single entity or as elements of a single thought. Interestingly, there is another aspect of GB, called *control theory*, which has to do with the occurrence and interpretation of a non-overt element, PRO.

⁷³ For an excellent and brief introduction to the main ideas, see Langacker (1990, chapter 1).

well). Unlike mental models accounts, the elements that are used to construct the complex conceptualization are *not* assumed to be objectively characterizable. Rather, many of the components will be structures resembling image-schemata (Lakoff, 1987; Johnson, 1987) or will rely on aspects of the cognizing process itself, as opposed to the objective scene putatively referred to.⁷⁴ I want to identify these 'conceptualizations' with emulator articulators (of course, there may be emulators, and perhaps even articulated ones, that will not enter into linguistic competence, e.g. low level musculo-skeletal emulators as discussed in Chapter Three). Emulators just are non-objective, action-driven models.

The final CG constructs are symbolic units, which consist of the association of phonetic units with conceptualizations. These associations facilitate the construction of conceptual complexes on the basis of sequences of recognized phonemes. An example might be that the phonological pattern 'both ... and ...' symbolizes the coordination (a cognitive operation) of the conceptual material symbolized by the two schematically characterized elements.

Consider the word 'on'.⁷⁵ Its core meaning (semantic pole) is a spatial relation in which one entity, a trajector, is above and in contact with another entity, a landmark. We might represent this configuration schematically as in Figure 6.8.⁷⁶ In the illustration the relation of being above and in contact with is represented by a thick line, which unfortunately precludes the actual contact in the illustration between the circles representing the trajector and the landmark.

⁷⁴ For a clear example of this, see Langacker (1990, chapter 2) in which it is argued that the ubiquitous Cora particles/morphemes for 'inside' and 'outside' rely not only on aspects of cognitive processing and perspective, but on recurrent patterns in the experience of the native speakers occasioned by the local topography and their activities.

⁷⁵ The example I will develop here is based on one developed in van Hoek (1992).

⁷⁶ We should note in passing that though CG makes use of such diagrams, they are to be understood as heuristics. Even something as seemingly straightforward as 'on' is highly context sensitive and non-objective. For example, 'on' makes implicit reference not simply to something being above and in contact, but more accurately, to something being in contact with a surface that is normally interacted with (something can be on the wall without being above it, but being on the table usually implies being on the upper surface).

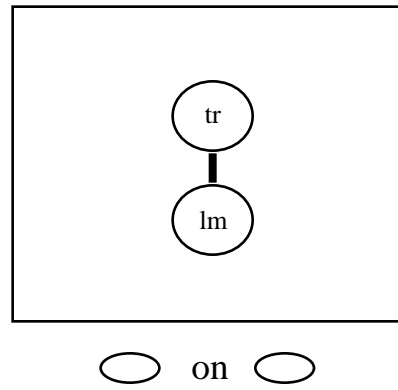


Figure 6.8: Schematic of 'on.'

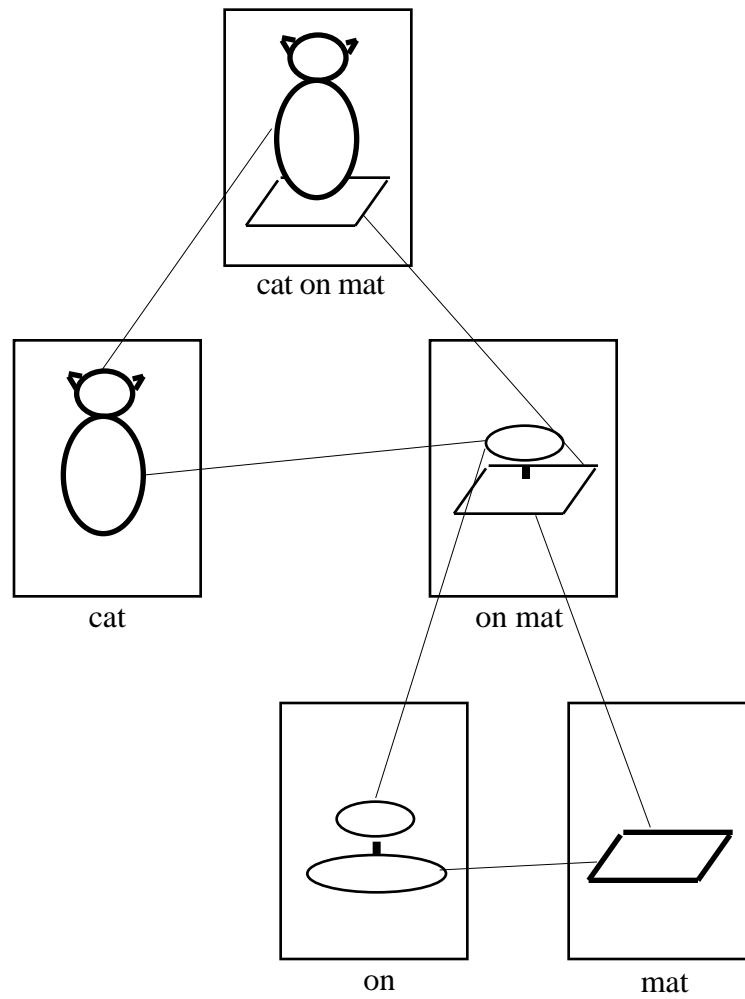


Figure 6.9: 'Cat on mat.'

'On' makes reference to two schematically characterized entities, the trajector and the landmark. In CG these are called elaboration sites, or e-sites. They are places where the component conceptualization links to another salient conceptualization, in a manner roughly analogous to the way functions accept arguments. We might elaborate the two e-sites with conceptualizations of a cat and a mat, respectively, as in Figure 6.9.

We can further specify this relation by construing this not simply as an atemporal spatial relation, but as a relation that is continuing through time, a sort of trivial process. This is the semantic contribution of the verb 'be' -- a schematic process that is unchanging and continuous throughout some span of time. This schematic process is elaborated by the relationship 'on', or in this case, CAT-ON-MAT, to form the conceptualization CAT-IS-ON-MAT (the contribution of the specifier 'the' is beyond the scope of this introduction, but see Langacker 1991)).

As a further illustration of CG,⁷⁷ consider the terms 'go', 'gone' and 'away.'

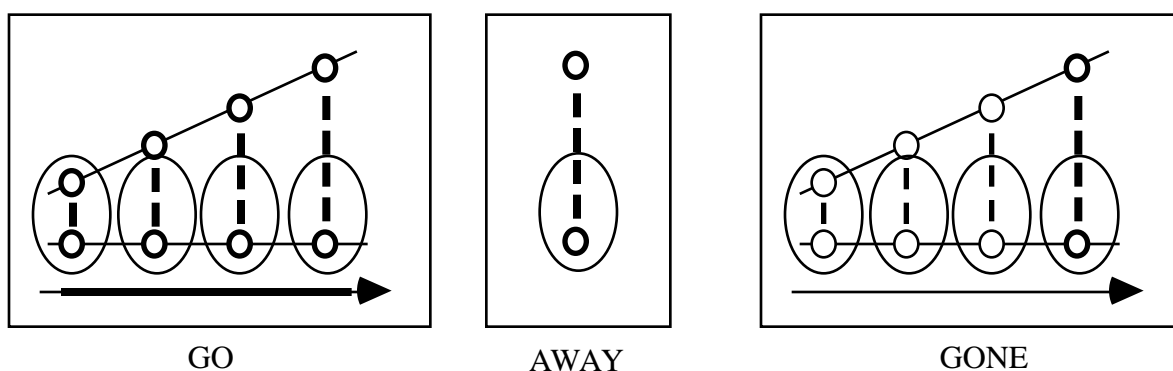


Figure 6.10: 'Go,' 'gone' and 'away.'

⁷⁷ This example adapted from Langacker (1990).

As this figure shows, 'go' is a process (hence the thick arrow indicating the passage of conceived time) by which a trajector (the upper small circle) moves out of the vicinity of some landmark (the lower small circle). By itself 'go' invokes *the entire process*. The participial morpheme '-en' (as in 'taken', 'stolen') has the effect of taking a process, detemporalizing it, and highlighting (or 'profiling') only the final state. When combined with 'go' to form 'gone', we have the conceptualization of a state which has come about as the result of some process, as in the right diagram. Thus the difference between 'go' and 'gone' is one of emphasis: go emphasizes the whole process, whereas gone invokes this whole process, but only profiles the final resultant state. By contrast, the term 'away' invokes the conception of a state in which the trajector is out of the vicinity of some landmark, as in the middle diagram of figure 6.10. Thus the difference between 'gone' and 'away' is that the former invokes the same relation, but as having come about by a process of going. By contrast, 'away' invokes only the spatial relationship, and is silent as to its history. The idea is that lexical items and morphemes operate by invoking conceptualizations, such as relations ('away'), processes ('go'), and cognitive operations ('-en', which turns a process into a relationship), etc. These will be the CG analog of emulator articulators.

Another crucial component of the semantic pole, according to CG, is the recognition of the importance of profile/base organization. Most, if not all, conceptualizations require that the profile of the conceptualization be specified in relation to some base. A clear example is 'hypotenuse', which is a line segment. But clearly 'line segment' is not enough to specify what a hypotenuse is. This line segment, which will be the *profile* of the conceptualization, must be understood as configured appropriately on the right *base*, in this case a right triangle. Similarly, 'tip' invokes as a profile a narrow or sharp entity, but as located on the base of an elongated object. 'Elbow' invokes a specific

junction, against the base of an upper arm and forearm. And so on. Grammatical consequences of this will be explored later in the chapter.

I take it as more or less obvious that Cognitive Grammar and ETM are fully compatible. CG maintains that semantics resides in cognitive constructions of certain sorts. This should be quite straightforward. Just as the capacity to imagine a scene or state of affairs is, as argued in Chapter Three, dependent on the use of some sort of reality emulator, so will the understanding of a description of a scene or state of affairs, as communicated by language. Furthermore, the component predications that CG invokes as the building blocks of complex conceptualizations will correspond to emulator articulants.

In addition, the profile/base phenomenon that CG makes use of is exactly parallel to the nature of the semi-locality which articulants were argued, in Chapter Four, to manifest. The example used, recall, was that of the articulants responsible, in the musculoskeletal emulator, for the forearm flexors. I argued that those flexor-articulants were essentially connected to certain other articulants in the musculoskeletal emulator, such as forearm extensors, hand proprioceptors, etc. This 'essential connection', however, is not to be confused with 'superposition.' Similarly, the profile is not to be confused with the base. It is an element of the base. It necessarily invokes the base for its own characterization, but it is independently targetable for all that.

6.5 Binding

There is a problem here, and it is a problem that touches not only sentence parsing, but cognition generally. It is the so-called *binding problem*. The problem arises when an explanation of the capacity to represent some objects or situations is couched in terms of component representations that must be structured into a composite representation. This is

clearly the case with the semantic pole of a linguistic expression as explicated in CG. It is also the case with certain sorts of emulators, as argued in the previous chapter, since the articulants of these emulators must be *re-structured* in the right way. In short, the price that must be paid by a representational system that takes advantage of systematic, productive components is some manner of structuring those components into coherent composite representations.

Consider the following toy vision example. Suppose that there are, in a certain visual domain, sixteen possible objects -- all combinations of the four shapes circle, square, triangle and oval, and the colors red, blue, green and yellow. One way to insure that a representational system can represent all such objects is to assign a distinct symbol to each such possible object. Thus one could imagine a system that had sixteen different symbols -- one for each of the objects it might encounter. Let us call this the *atomic solution*.

Another representational scheme might be called the *compositional solution*. In such a scheme each object is represented by a combination of features, such as RED SQUARE, or YELLOW TRIANGLE. This second scheme has some advantages over the first one. First, it provides an explanation of what, in representational terms, the red square and the red triangle have in common, namely the component RED. The atomic solution is unable to provide any such insight. Second, with the compositional solution, it is relatively cheap to expand the representational capacity of the system. If the domain gains a new shape, like a parallelogram, one need only add one new symbol to the inventory of the compositional system, whereas one must add (in this case) four new symbols to the atomic system.

It is well known that the primate visual system (as well as others) processes color and shape/motion information in separate pathways, and for purposes of the present conceptual point, we can assume that this is analogous to the compositional solution to the representation of objects. The problem faced by the compositional solution, and thus by the

primate visual system, is how to express the appropriate combinations to accurately represent the represented objects or scene. That is, suppose that the visual scene contains a red square and a blue circle. The compositional system will invoke the feature representations RED and SQUARE to represent the red square, and the component representations BLUE and CIRCLE to represent the blue circle. This would involve the color system representing the features RED and BLUE, and the shape system representing the features CIRCLE and SQUARE. But how does the system now distinguish the representation of a red square and a blue circle from the representation of a blue square and a red circle? Both will have exactly the same four components. What is needed is some way of representing the appropriate combinations. In other words, some way of *binding* the right features together.

One solution, championed by Fodor,⁷⁸ is to posit constituent structure in the language of thought. Whatever representational medium it is that has the capacity to represent red circles, blue squares, cats, being on mats, etc., makes use of a hierarchical tree structure to establish constituency relations among, ultimately, constituents that are compositionally atomic and context-independently tokenable. But this answer, whatever its other merits or faults, at best puts off the tough question, for now we must ask: How do *brains represent or make use of or implement* hierarchical constituent structure? Supposing that there is some brain state that represents blueness, and some other that represents circularity, *by what means* are they sometimes grouped together as a constituent, while at others, though they are active, they are not taken to form a constituent? Fodor's answer of hierarchical structure can seem less like an answer and more like a restatement of the problem. And so far, ETM and Cognitive Grammar are in the same boat as Fodor.

Before we address that problem, notice that something interesting has emerged from the discussion so far. The binding problem concerns not only linguistic

⁷⁸Fodor (1975).

representations, but perceptual ones as well. If the solution *in the case of language* is to posit a sort of syntax for representational combination, then this syntax cannot be language specific, unless one wishes a) to assume quite unparsimoniously that the brain solves the binding problem in two quite different ways in different systems, or b) to collapse all representation to linguistic representation (in which case, of course, the hypothesis of language specificity no longer has teeth). This latter possibility seems to be the path taken by Fodor. The Language of Thought is the representational medium not only for natural language, but for cognition generally. Whatever Fodor's convictions on the matter are, however, the possibility is now open that a solution to the *general* binding problem may point to a solution of the *linguistic* binding problem, and then to an account of constituent structure (or something equally powerful) generally. The rhetorically ideal situation would be one in which this *cognition-general* explanation of constituent structure can, by appeal only to mechanisms that underwrite it, shed light on *syntactic* phenomena which have only ad hoc explanations from within formal syntactic theory. I think that this ideal situation obtains. This will be the topic of section 6.7.

In Chapter Four I argued that the most refined and flexible emulators employed semi-local representations. These articulators, as I call them, are local in the sense that they can be independently targeted by cognitive processes, yet are not atomic (hence the 'semi-') in that they only have whatever meaning they have because of the (possibly flexible) contribution they make to a coherent emulational matrix. They are a profile which necessarily invokes a base for its characterization. If we assume that in the neural case such semi-local representations are tokened when the corresponding neurons are active, we buy for free the capacity to distinguish representations by the exact temporal pattern of this activity. In the simplest case, we can assume that representations are *phase* coded in

addition to being *frequency* coded.⁷⁹ For example, consider the following simple localist representations of the following features:

RED	[X]	SQUARE	[X]
BLUE	[X]	CIRCLE	[X]
GREEN	[]	TRIANGLE	[]
YELLOW	[]	OVAL	[]

Here the binding problem is manifest: Is the square red or blue? Let us imagine that these representations are expressed in neurons or groups of neurons (for simplicity assume a single neuron for each feature). Thus, out of the eight neurons, four happen to be firing strongly (i.e. high frequency), as indicated with the X's above. What is important to see is that neurons can have not only a frequency, but a *phase*. This phase can be exploited to bind the appropriate features by *phase-locking* those neurons whose representations are to be grouped. For example, supposing that the four active neurons are firing at a frequency of 100Hz (1 spike every 10ms), it might be the case that the RED and SQUARE neurons both fire at $t = 0\text{ms}, 10\text{ms}, 20\text{ms}, 30\text{ms}, \text{etc.}$, while the BLUE and CIRCLE neurons are both firing at $t = 5\text{ms}, 15\text{ms}, 25\text{ms}, 35\text{ms}, \text{etc.}$ Features may be bound into as many bundles as there are distinguishable phases. Notice that this possibility is not available to fully distributed⁸⁰ representations, because the representations for e.g. RED and BLUE would *share* representational vehicles, and would by definition always be phase-locked to each other.

There are interesting links between this solution to the binding problem and the management of (and mechanisms supporting) attention. Shastri (1993), for example,

⁷⁹ This chapter will discuss temporal coding in terms of phase-coding. This simplification is adopted for the sake of clarity. However, there are plenty of ways to exploit temporal structure beyond just the recognition of phase-locking.

⁸⁰ Here, as elsewhere, I use 'distributed' as a synonym for 'superposed'. This is perhaps not entirely conventional.

attempts to show that because of factors including membrane time constants, it is reasonable to expect that only a small number of phases could be distinguished reliably. He suggests that this might provide a basis for the limits of short-term memory, or alternately for the fact that attention can only be focused on a small number of distinct items at a given time. Wolf Singer (Singer (in press), Gray and Singer (1989)) has demonstrated, using recordings of cell activities in awake monkeys, that neurons responsive to features of visual stimuli to which the monkey is attending phase-lock, while these same neurons decorrelate (though, crucially, they do not stop firing) when the features they represent are within the visual field, but are not attended to. I will focus, however, on some simulations of sensory segmentation by Christoph von der Malsburg, and will here provide a brief characterization of his Dynamic Link Architecture (DLA),⁸¹ as it has been applied to sensory segmentation.

The model is given a 2-D representation of some visual scene. The task is to separate the scene, or segment it, into a figure and background. Each of the $N (= 36 \times 36)$ pixels of the image feeds to $M (= 8)$ feature units, each of which is selectively responsive to some feature or 'quality' which can be present in the scene (these would be the model's analogues of texture, shading, luminance, etc.). Upon presentation of the image, each unit starts spiking iff the feature to which it is sensitive is present at that pixel. The connectivity and dynamics of the model are such that it will settle into a pattern of activity in which there are two 'blobs' (collections of active units). Each blob is comprised of a number of units all of which fire in synchrony, and each of the two blobs is exactly (180 degrees) out of phase. The segmentation is shown to obey a number of the usual gestalt criteria for grouping. In short, what the model does is take its two distinguishable binding phases, and use them to bind two separate blobs of units, which can be interpreted as the figure and the ground.

⁸¹Cf. von der Malsburg and Buhmann (1992), von der Malsburg and Schneider (1986).

There are a few more points of interest, however. First, as von der Malsburg notes it would be entirely feasible for the DLA to be able to break the scene up into figure and ground in more than one way. Some sort of 'top down' signals, which biased the groupings in one way or another, could then effect a switch between the two groupings. Second, one can imagine an easy extension of the DLA that could process hierarchical structure. Von der Malsburg is worth quoting at length:

A more sophisticated system would sequentially subdivide the scene into a hierarchy of smaller and smaller segments, noting down the result of each step by abolishing connections between segments and by disambiguating situations with connections converging from several segments to a common target. In this way, temporal signal structure needs to express only those distinctions that are processed at a given time, previous distinctions being frozen into the connectivity structure. This would free the signals of no longer needed sets of neurons from having to be anticorrelated. Some disconnections can be permanently frozen into the system with the help of long term plasticity. Some disconnections, however, are valid only in the context of a given scene.
(von der Malsburg and Buhmann (1992))

We thus have the following picture of how the model segments a scene, and could hierarchically segment a complex scene. The model is able to distinguish a small number (two in this case) of distinct phases. It uses these phases to break the scene up into separate groups. Each group is bound by being phase-locked. The segmentation could then proceed to one of the blobs (objects, whatever) and segment *it*. But since there are only a small number of distinguishable phases, this can only be done by ignoring the segmentation of the larger scene, thus freeing the phases needed to break the segment into subsegments. If Singer (and others) are correct, then the portion of the scene that is currently being divided into phase-locked blobs constitutes the focus of attention. We will return to this capacity for hierarchical segmentation later.

6.6 Temporal characteristics of language comprehension

Though CG is quite powerful in its ability to handle a wide range of linguistic data, in most cases better than formal syntactic theories, I will not explore that power here. For present purposes, it will be enough to highlight the fact that according to CG, language use amounts to instructions for the configuration of complex conceptualizations. These conceptualizations are built from components that are interconnected via complex valence relations. In ETM terms, language use amounts to the configuration of emulators via the specification of emulator articulators. As the words of a sentence are heard or read, their valence features, together with conventional patterns of construction, yield a series of expectations and fulfillments. After hearing 'the cat is on,' a noun of some sort is expected, or perhaps a determiner and a noun. One has a conception built up with an open valence that schematically calls for some sort of thing. Or, in ETM terminology, one has articulated an emulator in which an expected articulator specification is missing.

Importantly, there is a structural mismatch between the acoustic signal used to code the phonological pole, and the semantic pole (and indeed the phonological pole itself). An acoustic segment of a spoken sentence can, qua sound, bear only a small number of simple relations to other acoustic segments, such as BEFORE, AFTER, and maybe quantified versions of these, like IMMEDIATELY BEFORE, etc. It is limited to a simple linear ordering.⁸² The phonological, and particularly the semantic structures, however, will typically be much more complex. A typical entity can bear a large number of distinct relations to other semantic entities. Consider:

⁸² This is not strictly true, since various aspects of prosody such as stress and intonation patterns provide structure which is superimposed on the linear phonetic sequence. I will ignore these subtleties, as none of the examples I discuss rely on such features (for the most part they will work as well if spoken in a mechanical monotone).

(50) After talking to his sister, John called his lawyer.

The entity 'John' entered into a communicative relationship (in this case a complex temporal relationship) with another entity. Furthermore, he already bears a kinship relation to this same entity. This communicative relationship, in addition, bears a temporal relationship (before) to another communicative interaction John entered into, namely calling his lawyer. And so on. Even this rather simple sentence demonstrates quite forcefully that the structure involved in conceiving a situation, semantic structure, can involve extremely complex relations assessed along a number of dimensions. The problem is how to code potentially very complex semantic structure using only a very limited acoustic structure. To provide a more concrete example:

(51) Sally painted the fence very quickly.

Semantically, the action 'painted' is associated with *three* separate constructs. First, there is the painter, the person who carried out the action. Then there is the fence, the entity which was the patient of the action. Finally there is the adverbial modifier 'very quickly', which provides more detail about the execution of the action. But notice that because of the restrictions imposed by linear acoustic order (i.e. an acoustic item can only be immediately next to *two* other acoustic items), 'very quickly', even though it is semantically associated with 'painted' and *not* with 'the fence', is next to 'the fence' and *not* next to 'painted'. Such order anomalies are more striking with extractions. In (52) the first word is most directly semantically linked with the last:

(52) What did Jerry and his friends think they saw?

The point is that because of this structural mismatch, important features of the semantic pole, like unelaborated e-sites, will often need to be kept in mind while other aspects of the structure are being processed. This is a real-time task which places demands on attention. To make a simplistic but nonetheless illuminating analogy, traditional formal linguistic theory treats syntax like a jigsaw puzzle. There are a number of pieces available, and they only fit together in certain configurations. If you change one of the pieces, other pieces must change to accommodate. And the problem is, given a set of pieces, how can one put them together, and how can one describe changes that can occur when one or more of the pieces are removed or change shape? But all the pieces are there in view, ready to be manipulated.

By contrast, I am arguing that linguistic competence is more like a game of Tetris.⁸³ One only hears one word at a time, and when each word is encountered, one must make decisions, very quick decisions, concerning the nature of the word, its valence relations, if it elaborates any previous e-sites, and if so which ones, etc. And each such decision provides more constraints on future decisions. The task is to be able to recognize opportunities for certain types of constructions as they arise -- an attention-intensive task. This is illustrated nicely with so-called garden path sentences, such as (53):

(53) The horse raced past the barn fell.

When the word 'raced' is encountered, most listeners fit the block 'raced' onto 'the horse' in the wrong way, a way which renders the rest of the puzzle insoluble.

6.7 What all of this buys us

⁸³ Tetris is a video game in which blocks of various shapes fall one at a time at a fixed rate down a two-dimensional arena until they hit bottom, at which point they are immobile. The player can manipulate the horizontal position and the orientation of the block as it is falling. The point, very roughly, is to get the blocks to fit together in such a way as to leave no gaps between them.

Before we get into the details of how the current picture of language can provide insight into the syntactic phenomena outlined earlier this chapter, it will be helpful to recap the major points.

A) Language works because brains associate phonetic (or graphic) material with conceptualizations (CG) or emulator articulants (ETM). Language provides a means for one person to configure semantic constructs in the head of another, and these semantic constructs will be emulators, i.e. non-objective models.

B) These articulants are only semi-atomic. Atomic insofar as they can be selectively targeted by cognitive processes (contra distributed representation), but non-atomic in that, since they have a function (/meaning⁸⁴) only because they have been articulated from some emulational matrix, they will necessarily invoke, or be invoked by, this matrix, or other of its articulants.

C) The construction of meanings is carried out by the interlinking of these articulants, which is a function of their own valence relations and conventional patterns of construction.

D) Structure within such complexes is coded and maintained via an attentionally mediated binding mechanism such as phase coding.

Scene segmentation and c-command

The suggestion that von der Malsburg makes for hierarchical segmentation would be formally identical to a binary branching tree structure, assuming two binding phases. In Figure 6.11, a two-dimensional shape is segmented, each of the segments are segmented, and so on. This is represented as a tree structure similar to the syntactic diagrams used earlier in this chapter. Segment F is broken up into blobs D and I. D can then be segmented

⁸⁴ In Chapter Seven I will argue for a form of functional role semantics.

into E and C, and so on. Let us now focus on blob A. A is encountered when C is segmented. It might be the *figure* of C, with B as the ground (or possibly vice versa). Or to adopt CG terminology, this bit of structure has C as the base, A as the profiled part of the base, and B as the non-profiled part of the base. Or in ETM terms, A is the articulant which is the focus of attention, B will be the articulant(s) which are invoked by A (e.g. A might be forearm flexors, B the forearm extensors and palm proprioceptive apparatus, and C the entire forearm).

So now let us imagine that the model (or brain) is zooming out from A to larger super-segments. That is, it fuses A and B to invoke C as a figure against the ground of E within the domain of D, and then fuses C and E to invoke D as figure against the ground of I within the domain of F.

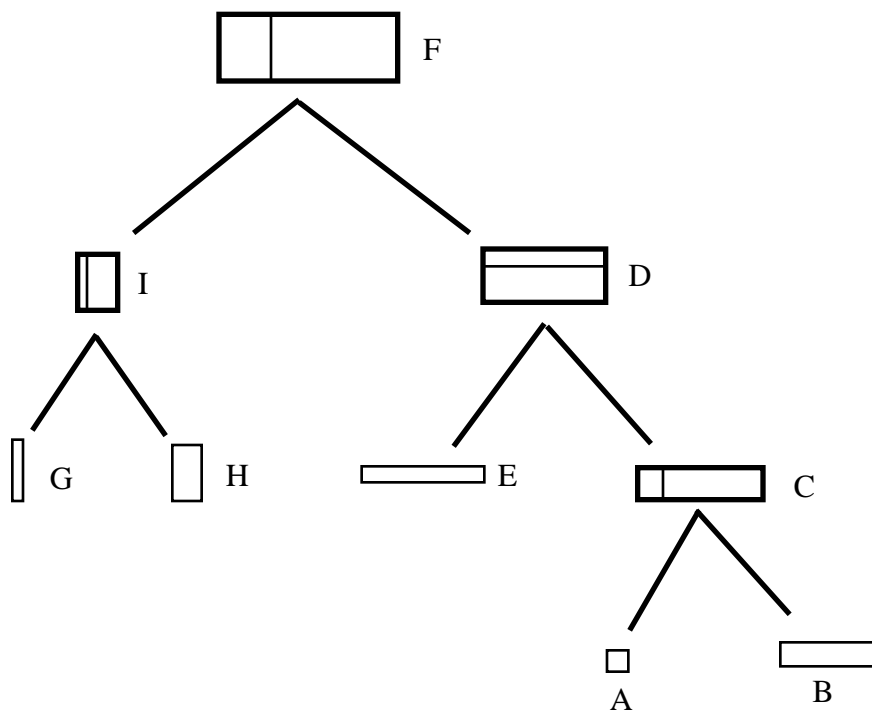


Figure 6.11: Hierarchical scene segmentation.

Interestingly, we can now characterize a visual-segmentation analogue of c-command. Recall that c-command was defined earlier in the chapter as (54):

(54) A c-commands B iff

- i) The first branching node dominating A dominates B, and
- ii) Neither A nor B dominates the other.

This is formally identical to:

(55) A c-commands B iff A is a *ground* recoverable from B by 'zooming out' zero or more levels.

The 'zoom out' path from A described a few paragraphs back invoked as 'grounds,' at the various levels, B, E and I, respectively. These are exactly the segments which c-command A according to definition (54). G does not c-command A because in order to recover G from A, one must not only zoom out to F, but then zoom in to G via I.

Von der Malsburg's model of perceptual segmentation thus can support what looks like a *syntactic* relationship between elements of the segmented scene. And if similar mechanisms are responsible for the binding of components of the semantic pole, we have what looks like a syntactic relationship defined over semantic elements. This 'syntactic relationship,' however, is not some separate process that takes place above and beyond the manipulation of the scene, but falls out of this *process of manipulation* given constraints on attention and binding. An internally emulated visual scene segmentation could plausibly be subject to the same constraints, which would follow if, as was argued in previous chapters, internal imagery and overt perception employ much the same neural substrate. And if this is the case, then components of this internally represented meaningful structure can be expected to have relationships which can be *described* as syntactic, but which are seen, rather, to be a shadow cast by the cognitive processes that manipulate the semantic structures.

Heavy NP-shift

Now let's turn Heavy NP-Shift, and the examples (23a) - (23d), repeated here as (56a) - (56d)

(56a) John introduced Mary to Linda.

(56b) *John introduced to Linda Mary.

(56c) John introduced to Linda the president of the company that just published her book.

(56d) ?John introduced the president of the company that just published her book to Linda.

According to the conventions of English, when the string 'John' is recognized first it can be taken to be the subject, or trajector (CG) of some relation or process that will be introduced later. So far no problem. The verb 'introduced' is a three-place predicate. One place can be linked to the already introduced 'John', and this leaves two other unsaturated e-sites (CG). The important point is that as soon as this verb is encountered, these two e-sites will demand separate free phases to bind the appropriate concepts (three actually, but the subject has already been bound to one of them). And because of the structural mismatch explained earlier, only one of these can be adjacent to the verb. It thus involves a processing cost to keep these two phases free. In the usual case with 'light' complements, we can assume that this cost is not too high. But if the first complement is long or complex, and has a degree of internal structure which requires the stacking and engaging of additional free phases, processing it can interfere with the brain's ability to keep the second phase free, and even to determine when to fill that phase. What one would like to do in such cases is to saturate the other e-site of 'introduce' quickly to establish that portion of the structure, and then go on to process the more complex component -- to prevent, in metaphorical terms, the brain from having to push the free phase stack too far. One still has the two phases in use, but one no longer has to expend attentional resources to keep one free while looking for its elaborator.

This solution is exactly what Heavy NP-Shift does. Presumably there is a processing cost associated with reversing the normal order of introduction of the complements, but there will be a point at which the maintenance of a free binding phase during interim processing outweighs that cost, and the reverse introduction, marked by the early occurrence of the preposition 'to' in this case, is implemented. All but one of the e-sites of 'introduced' is saturated early, and the attentionally mediated resource of free binding frequencies is freed for processing the final heavy NP.

Attention and extraction

As we have just seen, Heavy NP-Shift provides some evidence that the management of attention has grammatical consequences, and not simply as a performance spoiler. There is also impressive evidence that the processing of extractions is attention-intensive. The reasoning behind this is identical to the reasoning behind positing an attentional explanation for HNPS -- the phase of the neural populations representing the semantic structure associated with the extracted phrase must be kept free during the search for its proper binder (which will be in the neighborhood of the extraction site). We can thus expect that attention-grabbing or attention-intensive material that intervenes between the extracted phrase and the extraction site will interfere with the parsing of the sentence (in much the same way a heavy NP interferes with the ability to keep the second e-site open while looking for its elaborator).

Attention and Long Distance Extraction

Kluender (1990) argues that what he calls 'semantic barriers'⁸⁵ can interfere with the processing of extractions.

(57) How angry did Mary say that John was?

⁸⁵See also Ross (1987).

(58) ?How angry did Mary whisper that John was?

(59) *How angry did Mary editorialize that John was?

The gradation in acceptability here seems to be a function of the main verb. Kluender argues that open-class, referentially specific, low-frequency constituents are the hardest to extract over (and will conversely be the easiest to *extract*). Such items will naturally grab attention (compare the ease with which one reads right past 'say' in (57) as compared to 'editorialize' in (59)). This attention absorption is not a function of the processing of syntactic complexity, as in HNPS, but seems rather to be a function of semantic complexity. In trying to build up a conception, an internal model, in which one *says* something to someone, one can leave the communicative operation itself quite schematic. *Whispering* something, however, cannot simply be left as unspecified. The construction of the semantics of this is obviously more involved. *Editorializing*, furthermore, requires even more specifics and background for its proper understanding.

(60a) This is a paper that we really need to find someone to intimidate with.

(60b) This is a paper that we really need to find someone *we can* intimidate with.

(60c) This is a paper that we really need to find someone *that we can* intimidate with.

(60d) This is a paper that we really need to find someone *who* we can intimidate with.

(60e) This is the paper that we really need to find the linguist who we intimidated with.

(60f) This is the paper that we really need to razz the linguist who we intimidated with.

(60g) This is the paper that the audience really need to razz the linguist who we intimidated with.

(60h) This is the paper which the audience really need to razz the linguist who we intimidated with.

(60a) - (60h) are from Kluender (1990). Though making any hard grammaticality judgments is difficult, I agree with Kluender that there is a clear one-way gradation from good to bad here. In (b) the infinitive is replaced with a finite verb and the overt subject 'we' is made explicit. In (c) the complementizer 'that' becomes overt. In (d) 'who' replaces 'that'. In (e) the noun 'someone' becomes the referentially specific 'the linguist'. In (f) the high frequency and schematic 'find' is replaced by the low frequency more semantically complex 'razz'. In (g) the indexical 'we' is replaced by 'the audience'. And finally in (h) the wh-relative 'which' replaces the complementizer 'that'. Again, we are forced to conclude that semantics, and especially aspects of semantics that have an effect on the attempt to construct a *representation* of the situation described, have clear grammatical consequences.

Emulation, attention, and coordinate structure extractions

Given that the semantic structures that are the target of language use are emulators, and given that emulators are constructed from the agent's typical activities, then typical events, action sequences, and cause/effect sequences should be expected to play a more entrenched role in the operation of the emulator. It is easier for me to imagine (read: less attention-intensive) performing a lob serve than a corner kick, because I play a lot of racquet-ball, and very little soccer. And extended composite events such as going to the store and buying milk are more typical of my activities than going to the store and reviewing manuscripts. Though if my activities were different -- if I were a soccer player, or if I were in the habit of reviewing manuscripts in grocery stores -- then the details of my capacity to emulate these events would alter accordingly. Here we will find consequences for extraction from coordinate structures.

Lakoff (1986) and Deane (1991, 1992) have both examined the so-called across-the-board condition on extraction from coordinate structures. According to this condition,

extractions are licensed from coordinate structures only if the extracted element is present in, and then extracted from, every conjunct in the structure.

(61) Noam bought the book_i and gave it_i to Jerry.

(62) What_t did Noam buy t_i and give t_i to Jerry?

(63) *What_t did Noam buy the book and give t_i to Jerry?

(64) *What_t did Noam buy t_i and give it_i to Jerry?

There is, however, a class of exceptions to this constraint, in which an element can be extracted from only some of the conjuncts of what appear to be normal coordinate structures. What these exceptions have in common is that the conjuncts not undergoing the extraction have a clear semantic relationship to the other conjuncts, most centrally as providing details concerning the normal sequence of events, filling in elements of a script or frame evoked by conjuncts undergoing extraction, and the like.

(65) What_t did you go to the store and buy t_i?

(66) How much_t can you drink t_i and still stay sober?

(67) What_t do the guys in the Caucasus drink t_i and live to be 100?

Lakoff (1986) identifies three classes of exceptions to the constraint: natural sequences (as in (65)); violations of expectation (as in (66)); and cause/effect sequences (as in (67)).⁸⁶

The point here is that since emulators are constructed from normal interactions with the target system (the world in this case), such interactions and cause-event sequences will have privileged status over other less standard sequences. This privileged status within the emulator's operation is reflected grammatically as the capacity to conceive such situations as elements of a single coherent event sequence, as opposed to two separate sequences. So even though they seem to be, for all syntactical purposes, coordinate structures, they are semantically not really coordinate at all. Hence, the exceptional extraction.

⁸⁶Deane (1991, 1992) identifies a number of subcategories to, and extensions of, Lakoff's taxonomy. Furthermore, Lakoff, in an appendix, provides reasons for not treating these as parasitic gap structures.

Extraction from NPs

I argued in previous chapters that imagery is supported by emulators, and furthermore that the operations one can perform on the emulator-supported images are constrained by the operations one typically performs in the overt perceptual cases. One would not expect Mel's model of imagery described in Chapter Three to be able to rotate objects in a vertical plane, because it never moved around them that way, for example. The isochrony principle is another manifestation of this. If von der Malsburg's model of sensory segmentation is on the right track, and if perception and imagery do make use of similar neural mechanisms, then the fact that segmentation in the overt perceptual cases proceeds by breaking scenes up into segments, and then breaking those segments into subsegments, should also have its reflection in emulator-supported imagery. These facts should have grammatical consequences.

This appears to be the case. For example, when looking at a car, one would probably segment it into its most major components, which would be a body and tires, or maybe hood, trunk, tires, etc. The DLA model of von der Malsburg would work in this way. One would not expect the tire treads, for example, to be a salient segment of the car, though they might be a salient segment of the tires. This is perhaps because of simple attentional limits humans have. If we could distinguish more binding phases, tire treads and truck lock would be salient parts of cars. Given our limitations, however, we would expect to first segment the car into large parts, like hood, doors, tires, and then switch attention to the tires, or a tire, and segment *it*. We have here attentionally mediated scene segmentation in action. Consider the following sentences (68) - (72).

(68) Which hand_i did you burn your finger on t_i?

(69) ??Which arm_i did you burn your finger on t_i?

(70) Which tires_i did you like the treads on t_i?

(71) Which cars_i did you like the tires on t_i?

(72) ??Which cars_i did you like the treads on t_i?

Fingers are salient components of hands, but not salient components of arms. And treads are salient parts of tires, but not salient parts of cars. The only sense which can be made of the term 'salient' in these explanations is *cognitively* salient. The best sentences, by far, are those in which the matrix and extracted NPs have a figure/ground relationship. There appear to be two aspects to this. The first, exemplified in (69) and (72) above, is that skipping levels is bad. Extracting an NP referring to a salient part is acceptable (70), but extracting one referring to a salient part of a salient part is not (72). The DLA model would similarly have to first identify the appropriate blob in order to segment it into a figure and ground. The other feature is directionality:

- (73) I drew the hypotenuse on the small triangle.
 (74) I drew the triangle with the long hypotenuse.
 (75) Which triangle_i did you draw the hypotenuse on t_i?
 (76) *Which hypotenuse_i did you draw the triangle with t_i?

Clearly (75) and (76) are syntactically identical. Both involve the extraction of an NP from a PP modifier. The only significant difference seems to be that in (75) the extracted NP constitutes the base on which the matrix NP is profiled, while in (76) the profiled component is extracted from the base. The early invocation of the semantic pole of 'triangle' in (75) means that at that stage everything necessary for the characterization of 'hypotenuse' is already available. Or to put it another way, in order to invoke the image of a hypotenuse, one must first make available the notion of a triangle. If that notion is already present anyway, then it becomes easier to 'make mental contact' with the hypotenuse. This seems to imply that the cognitive operation of 'zooming in' is somehow easier, or less attention-intensive than 'zooming out'. Compare the ease of (77) as compared with (78).

- (77) The book is upstairs, in the bedroom, on the dresser next to the window.
 (78) The book is on the dresser next to the window in the bedroom upstairs.

Recall von der Malsburg's DLA model, and especially his suggestion for making hierarchical segmentations. It was that the segmentation proceeds by first isolating a segment, and then proceeding to divide it into subsegments. It would seem that, at least as far as the brain goes, this is a somewhat easier operation than starting with the subsegments, fusing them, and then zooming out to the higher level in which these fused subsegments form a segment.

6.8 Conclusion

The moral of the various phenomena covered in this section is that semantic and cognitive factors seem to play a key role in what sentences are and are not grammatical. Nonetheless, there are obviously a great number of linguistic phenomena that I have left completely untouched, and those I have touched I have done so only quite superficially. Rather, I have tried to show that at least in a few cases one can make some progress in addressing interesting phenomena with ETM, and how, in general terms, ETM can be part of an understanding of language use. And even though this has all the virtues of a post-dated check over cold hard cash, I think that the account here is interesting, and has promise, if pursued in more depth, to illuminate a wide range of linguistic facts without the gratuitous assumptions of language-specific representational formats. Post-dated checks, after all, do have their uses. At the *very* least, I hope to have shown the compatibility of ETM and Langacker's Cognitive Grammar. The latter *has* been applied to a fairly wide range of grammatical phenomena, and hence any significant compatibility would allow ETM to share, at least to some degree, the successes of Cognitive Grammar.

I am duty bound to make one final recognition of the debt this chapter has to a number of researchers. Most centrally to Langacker, but also to von der Malsburg, Deane,

Kluender, Singer, Kuno, van Hoek, Shastri, Llinas, and Crick and Koch. If any new heights are being reached, it is not because I am standing on the shoulders of giants, but because I have sketched a way in which they might stand on each other's.

Chapter Seven: Semantics

What the information is about - the reference of linguistic expressions - is not the real world, as in most semantic theories, but the world as construed by the speaker.

Ray Jackendoff

I have to this point been careful to avoid semantic issues. I have discussed aspects of the semantics of natural language in Chapter Six, but the treatment there at best forestalled the serious philosophical issues by not addressing them. That account was to treat natural language semantics in procedural terms,⁸⁷ as sets of instructions for constructing complexes of meaningful mental entities in the form of more or less fleshed-out emulators. But this account writes a large IOU: In virtue of what are these 'cognitive constructs' meaningful?

In the first sections of this chapter I will examine two sorts of answer to the problem of content: neo-Fregean theories according to which cognitive entities are meaningful in virtue of standing in some relation(s) to extra-linguistic (-cognitive) entities, and 'x-role' theories according to which meaning is a function of intra-linguistic (-cognitive) relations. The distinction is roughly the same as the one between externalist and internalist semantics. The treatment I will give to each of these candidate accounts of content will be schematic and incomplete. The point of the exercise will not be to embrace either view, but to get a feel for the terrain and to introduce the various intuitions,

⁸⁷ There are different uses that the term 'procedural semantics' has in semantics. One, which I don't mean to imply, is that the meaning of a term or proposition is the procedure or set of procedures for determining its truth conditions or reference conditions. Another, which I do mean to imply, is that terms are meaningful insofar as they are used as instructions, or procedures, for the construction of meaningful mental complexes.

conflicting requirements, and central issues that influence investigators to adopt, reject and modify the positions in the way they do. I will be satisfied if I can make explicit enough of these issues so that the motivation for a different synthesizing approach, which is as much a metaphysical enterprise as a semantical one, is clear. The emerging sketch will be a view that is (to use a suggestive metaphor) 'empirically' neo-Fregean, but 'transcendentally' x-role theoretic.

I hope to be able to provide accounts of various problems of intentionality, as well as show in outline how social, cultural, and other 'external' factors can play a role in determining content while not doing violence to the strong (and correct) Searlean intuitions that meanings are, in *some* important sense, intrinsically in the head.

7.1 Neo-Fregean Theories of Meaning

Frege (1952) identified the meaning of linguistic expressions with extra-linguistic entities. Names *mean* the object named: thus 'Frege' means Frege. Predicates, such as '... has a heart' also have as their meaning extra-linguistic entities, in this case functions from objects to truth values. Thus '... has a heart' will mean the function that maps all and only objects x_1, x_2, \dots, x_n onto True (where x_1, x_2, \dots, x_n are all and only the objects that have hearts).

Many current theories of content are naturalized versions of Frege's theory of meaning, and thus I will refer to them as 'neo-Fregean' theories.⁸⁸ These will include Fodor's (1987) Asymmetric Causal Dependence Theory, Millikan's (1984) Teleological

⁸⁸ There are, of course, many key differences between the theory of meaning constructed by Frege (1952) and those of Fodor (1987), Dretske (1979), etc. One such difference is what these philosophers take the extra-linguistic entities to be (for Frege: objects, functions, truth values, etc; for Fodor: properties capable of entering into nomic relations, etc.). Another difference is how they construe the relation in question (nomic, informational, or one of 'expressing'). Nonetheless, the important commonality I wish to stress by using the term 'neo-Fregean' is that these theories take the meaning of meaningful tokens to be, or be determined by, entities and objects external to and independent of the linguistic or cognitive system in question.

Indicator Semantics, and Dretske's (1979) Information Semantics. They are Fregean in spirit because they assign to the putative symbol an extra-cognitive⁸⁹ entity as its content or meaning. They are naturalistic in spirit because they all attempt to cash out the meaning-relation in naturalistic, typically causal-plus-some-bells-and-whistles, terms.

Frege's theory of meaning, however, is unable to deal with a broad range of psychological or intentional phenomena, and this led Frege to supplement it with a theory of sense.⁹⁰ For example, if 'Frege' means Frege, it also means the author of *Die Grundlagen der Arithmetik*, and the philosopher whose work Grush read on Monday. Similarly, '... has a heart' means, under benign assumptions, the same thing as '... has kidneys.' Thus 'Frege has a heart' and 'The philosopher whose work Grush read on Monday has kidneys' both mean exactly the same thing. Whatever the merits of this theory of meaning, providing a plausible account of understanding, or *psychological* content, is not one of them, as Frege clearly saw.

This problem, inherited by the neo-Fregeans, is the motivation for much of the subtle machinery hiding beneath the surface of their views. I propose to state the problem explicitly as a constraint on theories of content:

- (1) The content ascribed to a cognitive token by a putative theory of content must have the capacity to play a genuine role in psychological explanation.

Frege's attempt to handle (1) was his theory of sense. An expression, according to this view, has not only a meaning (= referent), but it 'presents' the referent in a particular way.

⁸⁹ Frege was, in the first instance, interested in the semantics of natural language, and so it is easiest in discussing Frege to refer to the underwriters of meaning as 'extra-linguistic.' This chapter will be concerned with the semantics not of natural language, but of cognitive or mental representations. Thus I will refer to their meanings as 'extra-cognitive.' Of course it will be possible to think about a thought (as well as to write about words), in which case the term 'extra-cognitive (-linguistic)' will not literally be true.

⁹⁰ Strictly speaking, Frege did not think that the theory of meaning, qua theory of meaning, was in any way deficient. The theory of sense was added to make sense of psychological phenomena. The culprit, for Frege, is not the theory of meaning, but our imperfect cognitive apparatus.

This 'mode of presentation' Frege called an expression's 'sense.' Exactly how one goes about explicating 'sense' in a more useful or concrete way is a difficult problem, one which we will be able to safely put aside. For now the following brief remarks will suffice.

Senses individuate contents according to the following criterion:

- (2) Two expressions, x and y , convey the same sense iff given a sentence P containing expression x , and sentence Q identical to P except that y is substituted for x , no rational speaker of the language could adopt different attitudes to P and Q (where adopting a different attitudes means to take them to have different truth values).

So, for example, 'Frege' and 'The philosopher whose work Grush read on Monday' have different senses, because a rational speaker might hold 'Frege was born in the 19th century' to be true, while being agnostic about 'The philosopher whose work Grush read on Monday was born in the 19th century.' Similarly a rational speaker might think that 'Bees have hearts' is true, while 'Bees have kidneys' is false.

In most normal cases, senses as individuated by (2) will do a good job of satisfying constraint (1). For example, if Smith says to Jones "The woman walking this way is a witch," we will do nicely by (1) if we assign the content '... is a witch' to the appropriate cognitive symbol, as opposed to '... is a persecuted woman.' Though in the appropriate circumstances we may suppose the two functors to express the same Fregean function (i.e. they map exactly the same entities onto True), the first, but not the second, will help us explain Jones' behavior, like fleeing madly and shouting "Gads! A witch! Run for your lives!" Such behavior would be difficult to explain if the psychologically relevant content were '... is a persecuted woman.' This may not hold up to more intense scrutiny. So be it. I do not mean to defend Frege's views, but merely to draw attention to them for illustrative

purposes. The present point is that the bare Fregean theory of *meaning* runs afoul of (1), for the simple reason that the world (which determines meaning) may not be the way the subject thinks it is (which is what determines behavior).

The neo-Fregeans who take the content of a symbol (type) to be the entity or property (type) that causes it (or covaries with it) must face this problem as well. So for example, in a move hardly less metaphysically extravagant than Frege's, Fodor takes it that an object may have as many properties as it takes to account for the different contents that object may support. Thus the legitimate properties that can enter into nomic relations with cognitive symbols, and thus underwrite their semantics, include the property of being virtuous,⁹¹ the property of being a unicorn,⁹² and in general at least one distinct property for every sense Frege would assign to account for (1).⁹³

One problem with neo-Fregean approaches to content is that in many cases it seems implausible as well as extravagant to posit the extra-cognitive objects and properties necessary to account for the full range of psychologically relevant contents. Thus, to take things to the extreme for a moment, if we are restricted to a neo-Fregean apparatus of

⁹¹ "All predicates express properties, and all properties are abstract. The semantics of the word 'virtuous,' for example, is determined by the nomic relation between the property of being a cause of tokens of that word and **the property of being virtuous**. It isn't interestingly different from the semantics of 'horse.'" Fodor (1990), p. 111, boldface emphasis added.

⁹² "I take it that there can be nomic relations among properties that aren't instantiated; so it can be true that **the property of being a unicorn** is nomologically linked with **the property of being a cause of 'unicorn's** *even if there aren't any unicorns*." Fodor (1990), pp. 100-101, boldface emphasis added, italic emphasis original.

⁹³ "... if the concept JOCASTA needs to be distinguished from the coextensive concept OEDIPUS'S MOTHER, that's alright because the two concepts are connected with (denote or express) *different properties*; viz., with **the property of being Jocasta** in the first case and **the property of being Oedipus's mother** in the second." Fodor (1987) p. 84, boldface emphasis added, italic emphasis original. Here, as in many places, Fodor's argumentative strategy is not clear. He presents this move as one possible way to get the appropriate distinctions of content from denotations, but he doesn't explicitly embrace it as *the way to go*.

In a different place, but using the same example, Fodor seems even more Fregean: "What's essential to my story is that believing is never an *unmediated* relation between a person and a proposition. In particular, nobody 'grasps' a proposition except insofar as he is appropriately related to a token of some vehicle that expresses the proposition... I can now tell you my story about Oedipus, which is that he had two different ways of relating to the proposition that J[ocasta] was eligible (and, *mutatis mutandis*, to its denial). One way was via tokens of some such vehicle as 'J is eligible' and the other way was via tokens of some such vehicle as 'O's M is eligible.' Fodor (1990) p. 167, original emphasis.

assigning extra-cognitive entities as the contents of cognitive symbols, we must have a type of extra-cognitive entity for every psychologically relevant content the cognitive system might employ. Furthermore, on pain of excessive parochialism and intolerance, we must assign entity types *indexed to individuals (or maybe cultures)*. So for example, if I look at a cooked beetle and think 'That's yucky,' and someone from another culture in the same circumstances thinks 'That's not yucky,' it seems we must really have two properties here, $yucky_1$ and $yucky_2$.⁹⁴ Also, to the degree that individual judgments differ, even within a language community, we must index properties appropriately, or else judge some set of preferences as correct and others as mistaken.

For example, if I respond to the Mona Lisa with utterances like 'That's not beautiful' and you respond with utterances like 'That's beautiful,' we have at least the following options: First, there is a real property, the property of being beautiful, and the Mona Lisa has (or doesn't have) this property, and one of us is just *wrong*. Second, there is a real property, the property of being beautiful, and the Mona Lisa has (or doesn't have) this property, and we are both right about this property (i.e. the relevant tokens in each of our cognitive systems either both have the content '... is beautiful' or both fail to have that content), but one of us is irrational. That is, even though the relevant token in your cognitive system has the content '... is beautiful' you utter "What an ugly painting" sincerely, refuse the painting when offered to you, etc.

The final possibility to mention is that there are two separate properties, the property of being beautiful_{me} and the property of being beautiful_{you}, the Mona Lisa has the first but lacks the second, we are both correct and rational. I don't think the neo-Fregean will like any of the three possibilities just explained. The first and second are excessively conservative and parochial. The last makes vast numbers of contentful states highly idiosyncratic -- nothing available to the neo-Fregean allows us to say that the property of

⁹⁴ We could, alternately, claim that the other individual (or culture) is just wrong -- beetles are yucky, period. I take it that this is not a live option.

being beautiful_{me} and the property of being beautiful_{you} have anything in common. They are different properties, and there is no more reason to call them species of the property of being beautiful *simpliciter* than there is for calling the property of being blue_{me} and the property of being beautiful_{you} both species of the property of being blue *simpliciter*, (unless we help ourselves to non-Fregean machinery, such as functional or conceptual role).

In light of this discussion, I think we can add the following constraint to our theory of content:

- (3) The content ascribed to a cognitive token by a putative theory of content must (in most cases) not be radically idiosyncratic.

What (3) is meant to do is to insure that some of the contents ascribed to tokens in one system will be (or can be) the same as contents ascribed to tokens in another system. Thus, roughly, tokens in my cognitive system having the content '... is beautiful' ought to have, at least roughly, the same content as the '...is beautiful' tokens in your cognitive system. By 'at least roughly' I mean, among other things, that the contents as ascribed ought to be able to support fruitful communication. The fact that we diverge in some cases in our judgments of beauty should not persuade us that we mean different things by '... is beautiful', nor should we conclude that we cannot communicate straightforwardly using sentences with words like 'beautiful' and 'beauty.'

It is interesting and instructive to notice how the tension between (1) and (3) influences thinking about content. (1) urges contents to be private enough to account for psychological explanation, and insofar as psychologically influenced behavior of different agents can differ importantly, we should recognize different contents. For example, '... is a witch' and '... is a persecuted woman' ought to be different contents, though perhaps elicited in myself and a 16th century Salem resident by the same objects. (3) on the other

hand urges contents to be more public and objective. We are, at least most of the time, talking and thinking about the same world, we communicate more or less effectively about it, and these facts ought to be explained by the type-identity, or at least type-non-idiosyncrasy, of many contents. I think one could explain a great deal of the lamentation and teeth-gnashing that is 20th century analytic philosophy of language as attempts to deal with this tension.

7.2 X-Role Semantics

Our next family of theories of meaning I will call x-role semantics. I mean to include here meaning holisms, functional role semantics, conceptual role semantics, causal role semantics, and the like. What I take to be the important commonality, and what contrasts maximally with the neo-Fregean views, is that on these accounts a term's (or symbol/state type's) meaning is a function of its relations to other such terms, symbols or states. Thus the meaning-bestowing relations are intra-linguistic or intra-cognitive, as opposed to views discussed in the previous section which took extra-linguistic/cognitive relations as semantically fundamental.

Thus, to express the point in 'language' mode, what makes a term in a language about cats is its position within a network of terms. It is because the term shows up in contexts like 'If Bubby is an x, Bubby is an animal', 'x's sleep most of the time', 'Look! The x is on the mat!' that x means *cat*. Of course, the other terms of these sentences are not given beforehand, and then used to determine the meaning of 'cat', but all such terms, articulating the interconnections of language as they do, rise to the status of meaningful as a whole.

The concern of this chapter, however, is the semantics not of terms of natural language, but of cognitive entities: brain states, cognitive symbols, etc. Roughly, it is the

role (causal, functional, conceptual, or whatever⁹⁵) that a given state type plays within a complex economy of such states that determines its meaning. So part of what will make cognitive state type X mean *cat* is that X is implicated appropriately with other states which mean *mouse*, or *dog*, or *animal*, or *pet*, etc. The point about holism will apply here as well -- these other states will mean what they do partly because of their commerce with X tokens. It is only against the background of a complex tapestry of such states and their dynamics that meanings can emerge. This contrasts clearly with the neo-Fregean theories, which assign meanings to a symbol without regard for its interactions with, or even the existence of, other such symbols.

Before I discuss some objections to x-role semantics, I'd like to examine how well, at a first pass, this approach satisfies (1) and (3). The neo-Fregean theories, recall, do a *prima facie* good job at handling (3). If the meanings of cognitive states just are the entities they represent or stand for, then assuming there is a single world we each interact with, our states cannot help but have similar, or identical, contents in the appropriate circumstances. The trouble for neo-Fregeans comes with (1), because (again, *prima facie*) what *counts* for determining or explaining my behavior is how I think the world is, and not how it actually is, and to the degree that these differ, the neo-Fregean meanings will fail to provide good psychological explanations.

X-role theories have the opposite fortunes. They cannot help, so it seems, but do well by (1). The meanings are assigned in such a way as to make it almost tautological that the contents ascribed will play useful roles in psychological explanation. But (3) has been the knock against such theories. If a state's role determines its meaning, then a different role entails a different meaning. Given the individual differences in cognitive dynamics brought about by different cognitive histories, genetics, etc., it seems unlikely that any state

⁹⁵ I will not feel obliged to make a stand on exactly which sort of role is in question because I do not *here* intend to defend any sort of x-role account (though I will later). The present section I take to be generic with regard to these theories, and the points and objections made to apply equally to all species.

in my head will have the same meaning as any state in your head, and communication and disagreement become sort of mysterious. If I think 'There's a cat' and you think 'That's not a cat,' our thoughts are not in conflict, because we mean different things by the state translated ambiguously as 'cat.' I'm **really** thinking 'There's a cat_{me}' and you're **really** thinking 'That's not a cat_{you}.'

Putnam's Twin Earth

Perhaps the most famous objection to functional role semantics, or any purely internalist semantics, is Putnam's (1975) Twin Earth objection. Suppose that, in a distant part of the universe, there is a planet almost exactly like present-day Earth, the single exception being that the chemical composition of their water analog is not H₂O, but a functionally similar compound XYZ. Suppose furthermore that on this planet, call it Twin Earth, there is an exact replica of me, call him Twin Rick. It will happen that the functional (conceptual, causal) dynamics of our respective minds are isomorphic, and since these roles are the sole determinants of meaning, the meanings of corresponding states should be identical. But they are not, because my water thoughts are about H₂O, while Twin Rick's are about XYZ. They thus differ in meaning, and so x-role cannot be all there is to meaning.

Burge's Arthritis

This example⁹⁶ is similar to Putnam's, and often interpreted as a simple generalization of Putnam's example to non-natural kinds.⁹⁷ The example is this: Assume

⁹⁶Burge (1979).

⁹⁷ Fodor (1987, p.29).

that I have a host of beliefs regarding arthritis, mostly correct beliefs, but at least one false one. Some of the former sort might be expressed as follows:

Grush believes that arthritis most severely affects the elderly.

Grush believes that his father suffers from arthritis.

Grush believes that his brother does not have arthritis.

And the false belief as:

Grush believes that he has arthritis in his thigh.

This final sentence might be true because I mistakenly believe that arthritis is not exclusively localized to joints, but can affect muscle tissue as well. We would express my false belief as above, even though I am unclear about the details of arthritis.

But now consider another duplicate of myself on another duplicate planet. Let's call him Triplet Rick to distinguish him from the twin in Putnam's example. Triplet Rick is molecule-for-molecule identical to me, and thus his cognitive economy is exactly isomorphic to mine. But assume the following difference between Triplet's world and mine: On Triplet's world, Triplet Earth (at least the Triplet English speaking part of it), the term 'arthritis', spelled and pronounced the same way, is used to refer to both arthritis as well as some muscular diseases (for clarity let us call this malady arthritis*). It so happens, then, that when Triplet Rick utters "I have arthritis in my thigh" he says something **true** -- *that he has arthritis* in his thigh*. However, we would not say that "Triplet Rick believes that he has arthritis in his thigh" because he has no (de dicto) beliefs concerning arthritis -- his beliefs concern arthritis*. Furthermore, and more interestingly, I have no arthritis* thoughts, but do have arthritis thoughts, **even though the structure of my notion of arthritis is closer to that of arthritis* than arthritis.**

The conclusion is that, even though our cognitive dynamics are the same, the symbol implicated in my thought I HAVE ARTHRITIS IN MY THIGH has the content *arthritis*, while the counterpart triplet-symbol does not. Therefore, x-role cannot be all there is to meaning. In particular, we are forced to acknowledge the role that the norms and practices of an agent's language community play in determining meaning, even when those norms have no obviously salient causal influence on my cognitive dynamics.

Multiple Assignments

Suppose, along with the proponent of x-role semantics, that there are a number of cognitive states, y_1, y_2, y_3, \dots , and that each of these bears various relations to other states as appropriate. The problem is that nothing about these relations determines a unique content for any of the states. Consider: though it may be possible, as in the example a few pages back, that state y_1 means 'cat', there is nothing within the holistic system itself that prevents the *entire system* from being mapped onto another domain (or differently onto the same domain) in such a way that y_1 gets interpreted as something else, like 'envy' or some natural number. In fact it can be shown that there will always be an infinite number of such alternate assignments.⁹⁸

Notice that this is not just an epistemological problem. The point is not that we would be unable to determine the 'real' meaning of y_1 , but that *there would be no determinant meaning*. X-role theories do not claim that the only way one can *discover* the meaning of a term or state is via some sort of holistic translation, but they make the stronger claim that the meaning is *constituted* by those holistically construed dynamic relations. If such systems do not determine a unique meaning, then there can be none (unless, of course, we help ourselves to some sort of content-fixing external relations,

⁹⁸ See Putnam, "Models and Reality," in Putnam (1983)

along the lines of some form of causal or neo-Fregean theory⁹⁹, i.e. give up x-role semantics).

Dretske's 'Sound and Fury' objection

It sounds like magic: signifying something by multiplying sound and fury. Unless you put cream in you won't get ice cream out no matter how fast you turn the crank or how sophisticated the "processing." The cream, in the case of a cognitive system, is the representational role of those elements over which computations are performed.

(Dretske 1983, p.88)

I think this is fallacious (you could get ice cream out if you put in a cow, some grass and maybe some water, the key being that cows are very complex in the right kind of way), but it does hit on a strong intuition. Why should it be the case that a symbol, meaningless in itself, should have a meaning in virtue of complex dynamic relations with other meaningless symbols? Searle's (1980) Chinese Room nurtures the same intuition -- make the processing and dynamics as complicated as you like -- you won't get meaning, even indeterminate meaning, for all that.

7.3 Towards a positive theory

I think that one lesson to be learned from the above considerations is that we must be careful to distinguish two aspects of meaning. Accordingly, I plan now to sharpen some dull terminology.¹⁰⁰ Leaving the term 'meaning' as a generic undifferentiated term, let's

⁹⁹ Putnam intends this argument to tell equally against this possibility as well, his point being that in order to use causal relations to fix content, one must assume that one has pre-theoretic knowledge of what causal relations there *really are*. Putnam takes it that these causal relations are as much open to reinterpretation as the referents themselves. I think Putnam is right, but for now the weaker argument against x-role theories will suffice.

¹⁰⁰ The terminology used by semanticists seems to be largely idiosyncratic, to the point of making

distinguish 'content' from 'reference' in the following way. Let content be whatever it is about a token that allows it to play an explanatorily robust psychological role. It will be the personal, solipsistic aspect of meaning. And let us take 'reference' to be the real-world entities, properties, states of affairs, etc., that cognitive states signify.

This is still rough, but we're a lot better off than we were before. For example, consider Dretske's 'Sound and Fury' objection. Though it hits on one intuition, it beats up on another. Suppose that I am, in fact, a brain in a vat. It will turn out that, because my brain enters into none of the appropriate informational relations, I have no 'representational' states according to informational theories of representation. My brain really is just multiplying sound and fury, and signifying nothing, on this account. Though I don't wish to squabble over who gets to use the terms 'representational' or 'signify', there is something that my brain *would be* producing -- at worst a sort of 'mock significance.' That is, it certainly seems to me, brain in a vat (or in a skull), that I have thoughts with some sort of content and meaning. It may not be *representational*, but it's not a nothing, either. Well, we can just call whatever it is about those cognitive states that makes them have the meaning they seem to for me their *content*. We can agree (though we don't have to) that these states don't have any reference, or at least not their 'usual' reference.

I take it that the Twin Earth case can be handled the same way. Even though my water thoughts and my Twin's XYZ thoughts have different referents, they do have something in common: these thoughts will have exactly isomorphic effects on our cognitive system. This commonality, whatever it amounts to, will be their content. The problem is that the content that the brain in the vat has, and that I and my Twin share, seems upon reflection to be an odd beast. It doesn't really mean water, H₂O or XYZ. These might, in appropriate contexts, be the *referents* of these meaningful states, but they cannot be the contents. Furthermore, neither I nor my Triplet has states with the content *arthritis*. We do

communication difficult in many cases. The way I plan to sharpen the terminology is thus by no means canonical, as there is no canonical terminology so far as I can tell.

have states with the same content, and whatever that content is, it is such that on Earth it expresses arthritis, and on Triplet Earth it expresses something else (arthritis*).

One is tempted to join Fodor (1987, 1990) in describing content as a function from contexts to meanings,¹⁰¹ or to join Block (1987) in advocating a two-component theory,¹⁰² x-role theory plus some 'reference fixing' component, which could take the form of some causal theory of reference, and/or something along Burgean lines. I propose, however, that we hold off on this. We have examined some of the more well-known and influential positions on content, as well as some of the intuitions behind them and against them. While the treatment has been somewhat rough, has blurred some distinctions, and has ignored many considerations, I think it has been adequate to its purpose, which is to establish some terminology and provide a background of motivations and concerns against which I will begin to sketch an alternative.

7.4 Interpretational Semantics (IS)

We are now in a position to start saying some positive things about meaning. I want to start with an account put forward by Robert Cummins (1989), which he calls Interpretational Semantics. Though there is a lot that I take to be wrong about Cummins' view (as will be elaborated presently), at least two aspects of it are correct. First, it is at heart an x-role account, and second, it has the capacity to explain the correct relation between content and reference (though Cummins himself doesn't make anything of this).

First, I will present Cummins' Interpretational Semantics (IS). The presentation here will be fairly quick, since the ideas are quite clear. Next, I will bring some objections

¹⁰¹ Fodor makes a distinction between narrow and wide content. Narrow content corresponds roughly to what I am calling content. The similarities end here. Fodor's idea is that given a narrow content, and a context, one can determine meanings, or wide contents.

¹⁰² Block wants to maintain a version of x-role theory, but claims that some of the terms have their meanings fixed via relations to extra-linguistic entities. His view is thus a hybrid of the two I have been discussing.

against this view, many of which were discussed in connection with x-role theories earlier. I hope to show that bare IS does not fare very well in the face of these objections. Then the overhaul will begin. Though IS provides the correct sketch, little but its rough form will be discernible in the finished portrait. A better analogy: IS will be shown to be a limiting case - like the ideal gas law or Newtonian Mechanics, a useful heuristic in many cases, but literally false. Thus, in the same way that the ideal gas law is, in standard physics and chemistry texts, first introduced, and then gradually improved by taking account of, e.g., the fact that molecules take up space, I will attempt to identify the idealizing assumptions of IS, and then begin relaxing them.

Following the construction of the theory, which I will call General Interpretational Semantic Theory (GIST), the remainder of this chapter will take on some further tasks. The first will be to show exactly how GIST fits into, or is appropriate for, ETM. This will involve showing the relationship between emulation and interpretation. Finally, I shall show how GIST/ETM deals with some traditionally recalcitrant phenomena surrounding intentional (or opaque) content attribution.

What follows is more or less the view expressed by Robert Cummins (1989). Let us start with a straightforward example, an adding machine. Assume, for simplicity, that the machine in question is just an *adding* machine, i.e. it cannot multiply or subtract directly. Commonsensically, such a device adds -- it accepts numbers as inputs and produces sums as outputs. But this common-sense interpretation isn't literally true, though. Numbers are, at best, abstract entities. Clearly mechanical devices don't *produce* or *interact with* numbers! In fact, adding machines don't directly interact with numerals, either. We are better advised, I think, to talk in terms of sequences of button-pressings and display states.

But this is to chuck the baby with the bath water. We no longer have, it seems, an adding machine, but a complex physical object, one which is structured such that tokens of

certain input-state types (button-pressing sequences) reliably produce tokens of certain display-state types. This may be interesting, but it is not clear what licenses its being an *adding machine*. The world is full of interesting physical systems that produce reliable state transitions, but are not, *prima facie*, adding machines.

Well, maybe the fact that the buttons are labeled with numerals and the outputs states are numerals as well is what makes the difference. The fact that this is the case makes it possible to interpret the input/output transitions as instances of addition. That is, the reason the mechanical device is an adding machine is that it is physically structured in such a way that button-pressing sequences, *interpreted* customarily as numerals in accord with their labels, produce output states that are also *interpretable* as numerals, and furthermore, the sets of inputs and outputs thus produced satisfy the addition function.

Enter Cummins' 'Tower Bridge' diagram (see Figure 7.1). The lower span is the physical/causal span. The material on the left side of the lower horizontal arrow represents the initial (input) states, the sequence of button pressings, for example. The arrow is a causal arrow, indicating that the machine is structured such that being in certain physical states, input states, causes it to go into others, output states. The latter are indicated on the right side of the lower arrow.

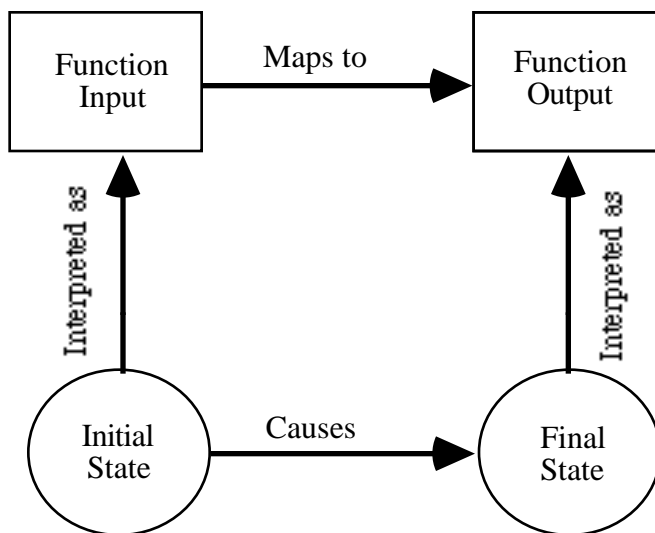


Figure 7.1: Cummins' Tower Bridge.

The upper span is the mathematical span. It represents the function (or in general, the target domain) that maps pairs of numbers onto numbers, the former being on the left side of the arrow, the latter on the right. The arrow itself is taken to represent the mapping function (addition in this case). The vertical arrows are interpretation arrows. When going up, they mean roughly "is interpreted as," and going down "is the interpretation of." So, we see that the initial state sequence of the mechanism gets interpreted as an ordered pair of numbers, while the final state gets interpreted as a number as well.

Finally, we can say why the mechanism is an adding machine. The short answer: Because we can (and do) easily *interpret* it as an adding machine. Longer answer: Because there exists a consistent¹⁰³ interpretation of the relevant machine states under which the causal transitions of those states satisfy the addition function.

We are not restricted to explaining the semantics of adding machines, of course. Given a target domain with some amount of structure, i.e. identifiable entities and identifiable relations between them in the domain, and some candidate 'cognizer' of that domain, we can say that if there is a consistent interpretation of the states of the candidate cognizer, according to which those states and their relations mirror the structure of the target domain, then the states (and relations) in the candidate cognizer are *about*, or *represent*, their corresponding counterparts in the target domain, under that interpretation.

Target domains might be anything: addition, my right arm, the world. Thus the potential applicability of IS to ETM should be obvious: emulators *emulate* some target domain. IS provides the promise of accounting for the semantic properties of the articulators

¹⁰³ We must assume that there is a way of identifying state types that is useful for the interpretation, and such that this identification is invariant across different possible state transitions. To see why this is important, consider a mechanism that exhibits at least the following two transitions: first, when I press the button labeled '2' and then the button labeled '2' again, its display state outputs a '4', and second, when I press a '3' followed by a '1' it outputs a '4'. Unless we identify the types in some invariant manner, it will be possible to interpret this as anything, such as a multiplier, provided we interpret the second occurrence of the '4' in the output as representing the number *three*.

of emulators in the most straightforward way possible: An articulant is about (/represents/refers to) the entity(ies) of the target domain that is its counterpart under the interpretation that licenses the emulator being a real emulator in the first place.¹⁰⁴

The feeling that this is too easy, that you have been hoodwinked, is entirely justified. There are a lot of problems with this account, some of which seem fatal. But I think we can fix it.

¹⁰⁴ This is a good place to bring attention to the essential role played by the representational format argued for in Chapter Four. Semantic evaluation on this theory requires that the emulator operate via the interaction of independently manipulable articulants. Standard connectionist orthodoxy is that such entities are fictitious and have no identifiable incarnation at the implementation level. Such assumptions make semantic evaluation nearly impossible (hence the close tie between connectionism and eliminativism), and this makes such devices as cluster analysis and principal component analysis, however genuinely unenlightening they really are, among the only straws to grasp. Two points: First, there is no incompatibility between the notion of an independently manipulable articulant and the spirit of connectionism. Second, such articulants do not imply localist coding or grandmother cells. What is required is that at any given moment, the vehicle for representing one articulant does not support any other representation. But such episodes could be handled by a resettable workspace or blackboard that is used to implement only one model at a time.

7.5 Problems with Interpretational Semantics

According to IS, what allows one to interpret a state as having a particular content is that that state and its x-role (which just is the causal role the state plays on the lower span of the bridge in this case) is isomorphic to some entity and its relations in the target domain. Thus, in addition to whatever problems beset the interpretive aspect of the theory, it will inherit all the problems of x-role accounts as well. Most of these were mentioned in earlier sections of this chapter, but I will briefly re-list them here, and add a few more.

Alternate Assignment Problem: In general, any given cognitive system will have a structure that is isomorphic to many different target domains.

Putnam's Objection: I and my Twin have the same cognitive architecture, so why is it that I have water thoughts and he has XYZ thoughts?

Burge's Objection: I and my Triplet have identical cognitive structures, so why do I have false arthritis thoughts and he true arthritis* thoughts?

IS too conservative: In order to be interpretable as representing the world, a cognitive structure must be isomorphic to the world's structure. If the isomorphism fails, which it must for all non-omniscient beings, then we cannot really represent the world.

IS too liberal: Lots of physical entities have structure, and there are lots of potential target domains, so how can we avoid the conclusion that any given physical entity, under some interpretation, represents all sorts of wild things?

IS fails as reductive account: By invoking the notion of an interpreter, IS ultimately explains semantics in terms which themselves are semantic, on the plausible assumption that interpretation is a semantic phenomenon. Thus IS answers Dretske's 'Sound and Fury' objection, but at the price, so it seems, of circularity.

And just for completeness I will include here the two primary constraints developed earlier in this chapter, even though they are not necessarily independent of the foregoing objections:

- (1) The content ascribed to a cognitive token by a putative theory of content must have the capacity to play a genuine role in psychological explanation.
- (3) The content ascribed to a cognitive token by a putative theory of content must (in most cases) not be radically idiosyncratic.

My task is to modify and extend IS in such a way that it answers these objections and satisfies the constraints. I think that the theory that we will end up with is both correct and enlightening. Unfortunately, there will be loose ends and unaddressed issues brought about by limits on space and on my ability. Nonetheless, I am confident that the charitable reader will share my belief that GIST is a promising account.

7.6 The General Interpretational Semantic Theory (GIST)

IS, at heart, explains semantics in terms of three other things: the (candidate) emulator (the lower span and its causal structure, in Cummins' parlance), the target domain (the upper span), and the interpreter. Furthermore, though I don't think Cummins would like to think about it this way, it really puts the bulk of the semantic work in the hands of the interpreter.¹⁰⁵ The bottom line is that 'cat' means CAT because I (or someone)

¹⁰⁵ This truth is not at odds with IS's being vulnerable to the objections to x-role theories, as explained in the previous section. We simply rephrase the objections with the interpretive terminology. For example, the alternate interpretation problem.

interprets it that way. The rest of the story is best viewed as placing constraints on what one can and cannot legitimately interpret as what, namely, that a given interpretation is licensed only when the appropriate isomorphisms of structure exist. These 'isomorphisms of structure' are relations between not just the symbol and thing symbolized, but between the entire emulator and the entire target domain. Hence, holism.

This core idea, that semantics is a matter of constrained interpretation, I think is correct. This much I share not only with Cummins, but with many others as well. Dennett (1987), for example, holds this core idea. The constraints on interpretation are roughly pragmatic constraints for Dennett, which single out certain sorts of interpretive stance as appropriate on a given occasion. For other famous accounts, the constraints take the form of principles of rationality, or charity, or humanity, etc. The list is a familiar one. For Cummins, again, the constraints take the form of structural isomorphisms.

7.6.1 Target Domains

The reason I chose to use Cummins' theory as a starting point is that, unlike Dennett (1987) and Davidson (1973), for example, Cummins makes explicit use of the notion of a target domain (though his terminology differs from mine), and his theory is more readily applicable to ETM, because IS depends on isomorphisms between (brain) states and target domain entities (as opposed to, e.g., translations between instances of language use).

Let's begin with Cummins' use of the notion of a target domain. This is useful because it gives us an easy way to characterize the difference, introduced earlier in this chapter, between content and reference. The *content*, on the IS account, will just be the place occupied by a physical symbol in a causal (or functional, etc.) nexus of other such symbols (how this gets characterized will be addressed presently). The *reference* will be that entity of the *target domain* which corresponds to that state under an interpretation

(reference is thus interpretation dependent in a way that content is not). Figure 7.2 provides a generalization of Cummins' Tower Bridge. Here, the circles on the bottom, labeled 1 - 5, represent emulator articulators, and the thick lines connecting them represent relations which obtain between them. The squares A - E on the upper span represent entities in the target domain. The thin vertical lines represent interpretive relations, as in the tower bridge.

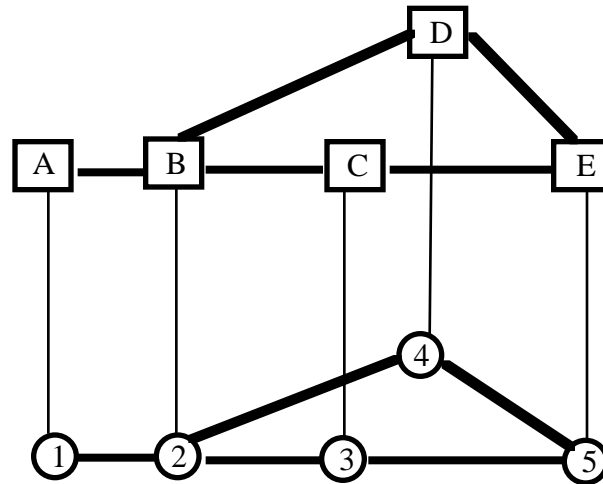


Figure 7.2: Generalized tower bridge.

What generally happens is that since x-roles (and hence IS content) are such difficult things to get a handle on (how does one conveniently describe or understand the details of such a sophisticated causal structure?), people characterize the *content* by means of the *referent*. That is, in order to talk about the content of a certain cognitive state, such as '2,' it is convenient to invoke its referent, in this case B. This is convenient because when making such an attribution, we presumably are already familiar with the structure of the world (or realm of reference) and we have antecedently decided to interpret the emulator in terms of the target.

This strategy is quite useful. I like to think of it as using the referent as a handle on the content (later in this chapter this idea will be sharpened). When the communicating or

attributing parties are in agreement as to the nature of the realm of reference as well as in agreement concerning the isomorphisms between the representing system and this realm, then the use of reference handles seems so natural that it can lead us to think that *content is reference*. This in turn leads to some of the confusion within semantic theorizing. The problem with *equating* content and reference is that such contents fail to be psychologically adequate when the world is not the way the cognizer thinks it is. Thus, on the IS account, when structural isomorphism fails we have two choices: Either insist that the 'referent'-characterized content **really is** the content of the state (prime facie differences notwithstanding), or insist that since the exact isomorphisms of structure do not exist (because the cognizer has some errant ideas about the world), the putative cognizer is not really a cognizer.

For instance, suppose I am trying to ascertain the meaning of some of Democritus' cognitive states. There will be certain states in his cognitive system that might be characterized with the content '... is an atom.' The problem is that this state will occupy a location in a cognitive complex which is not exactly isomorphic to the structure of the real world (because, e.g., Democritus' atoms cannot be split). Thus *either* we can insist that the state in question really has the content '... is an atom,' where 'atom' here means roughly what we, or current-day physicists, mean by it. This assigns a content which runs afoul of constraint (4), since if Democritus were asked if atoms are composed of smaller parts, he would presumably have answered that they are not. *Or*, we could insist that Democritus does not represent the world at all, i.e. is not cognitive. If IS requires an exact isomorphism between target domain structure and cognitive structure, then Democritus' brain does not support a cognitive structure (at least not one that represents the world), since we cannot interpret it as such.

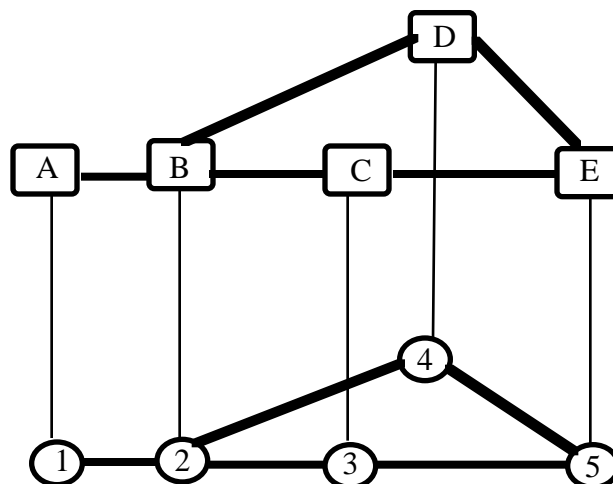


Figure 7.3: Tower bridge with non-real target domain.

The first objection is just the standard objection raised against the neo-Fregeans about their failure to handle (1). The second is the conservatism objection raised in the previous section. There is an easy answer to these problems, which is the first modification I propose to make to IS. Following the advice of Jackendoff in the quote at the head of this chapter, we should construe 'the target domain' more liberally, to include not (just) the actual world, whatever that may be, but the world(s) of the cognizer in question. So, although the structure of Democritus' cognitive system does not match the structure of *my* world (the real world?), it *does* match, indeed it constitutes, the structure of *his* world. What we must do in order to understand Democritus, to assign content to his thoughts, is to get into his head, into his world. We construct a model, or emulator, that we feel matches the world of Democritus. Entities in this world now become the reference-handles. Figure 7.3 is quite similar to Figure 7.2, except that the squares have been replaced by rounded squares, indicating that they are entities not in the real world (or target domain), but in the world of the cognizer in question. Thus we can now make the following claim: The content of a state will be its referent in the 'world' created by the representing system,

the emulator. The reference of a state will be the entity in the real world with which it is isomorphic. Thus Democritus' cognitive tokens of ATOM had the *content* (roughly) of 'indivisible material particle', but *referred* to (if anything) the divisible atoms of current physical theory.

Before I expand on this thought, I want to answer an objection. I have loosened the requirements on the target domain for a candidate cognitive system from *the* world to *a* world, but have given no idea of what can and cannot count as a potential world. Haven't I just flattened the notion of representation out so thin that everything trivially represents its own little world, no matter how simple? Take a system, any system. It will have a certain amount of structure, perhaps not much, but some. If we are free to make worlds willy-nilly, then each such system will 'have a world' for it, defined just exactly to have the structure of the system in question. Surely this is madness. My first response to this is that I am using the term 'potential world' in a way that is not meant to let anything go. I won't try to offer any hard-line conditions or definitions. I'll rest content with the thought that just as questions of content are ultimately matters of interpretation, so ultimately are questions of personhood and of representational status generally. Normal humans represent a world to themselves, and most likely most primates do as well. Single-celled creatures probably do not.

But I *can* say a few things. First, a world is something that is independent of the organism in it. Thus, perhaps, one of the more important criteria in deciding if some creature represents a world is whether it seems to internally emulate external features of the world in a way that is, to some degree, independent of immediate sensory stimulation (i.e. not in a closed control loop -- see Figure 3.1). Furthermore, a world must be *in principle* understandable by us *as a potential world*. I can, to varying degrees, construct mentally (i.e. commission a sub-emulator for) the worlds of Democritus, of an ape, of a child, of a member of an exotic culture, maybe even a bat (to a very limited degree). But a clock? a

television? a Cray II computer running Word Perfect (very quickly)? Structure, even very complex structure, is not equivalent to world-like structure.

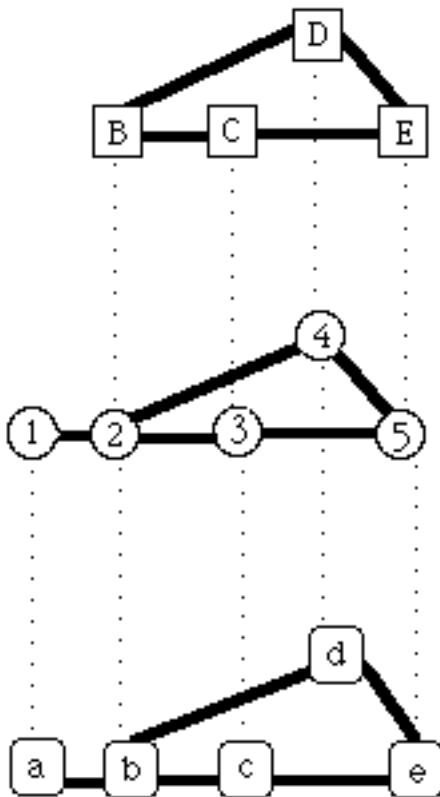


Figure 7.4: Two interpretive projects.

The cognitive structure (center), containing the elements 1 - 5 represented by circles, is subjected to two different evaluative projects. The first (upward interpretation) maps this cognitive structure onto the 'real' target domain (containing entities B - E, represented in diagram by squares) to determine *reference*. Representation (1) has no real world counterpart, i.e. does not refer. The bottom structure represents the cognizer's world, containing entities a - e, represented by rounded squares. This mapping shows the *content* of the representations, thus (1), though it does not refer, has content (a). This might be, e.g., the Great Pumpkin or a UFO.

Interestingly, I think that many of the constraints or 'interpretive principles' that have been championed in recent decades fall out naturally as conditions on our ability to

easily construct world-like sub-emulators. We are not very good, perhaps, at imagining worlds with objects that routinely have and lack a given property at a given time, for example. This constraint manifests itself as a pressure to interpret others as *rational*. We are understandably better at imagining worlds that are similar to our world, i.e. to what we might wish to call a human world, and this induces us to interpret others as being like *ourselves*. Thus the appeal of the principle of humanity.

We have taken our first step from IS to GIST, which is simply to liberalize our idea of what counts as a legitimate target domain, to recognize that different cognitive entities represent the world differently, or more accurately, represent different worlds to themselves. This modification shows how we can handle the charge of parochialism leveled at IS, and how GIST content is adequate to constraint (1). Indeed, given some constraints supplied by limitations on what qualifies as a world, we can address the charge of excessive liberalism as well. But there are additional benefits here. First, we can now characterize the difference between content and reference more adequately. Content, now, will be the interpretation of the cognitive state *in the cognizer's world*, while reference is its interpretation in the real world (see Figure 7.4).¹⁰⁶ Some of the problems about reference, such as its inscrutability, are then shown to be by-products of the fact that there may not be the appropriate structural isomorphisms between the cognizer's world-emulator and the real world (or my world-emulator). The adequacy of an interpretation will be a matter of degree.

7.6.2 Constraints on interpretation (part one): Putnam

So far we have articulated GIST enough to address the problems of liberalism and parochialism (relying on the notion of a 'world'), to characterize, if roughly, the difference between content and reference, and to satisfy constraint (1). In this section I will address

¹⁰⁶ What counts as the 'real' world is an issue which I will not be able to address. For those who get trigger-happy around realists, substitute 'the interpreter's' for 'real'.

the Putnam and Burge objections. We will see that the solutions to these objections will also take the form of interpretive constraints.

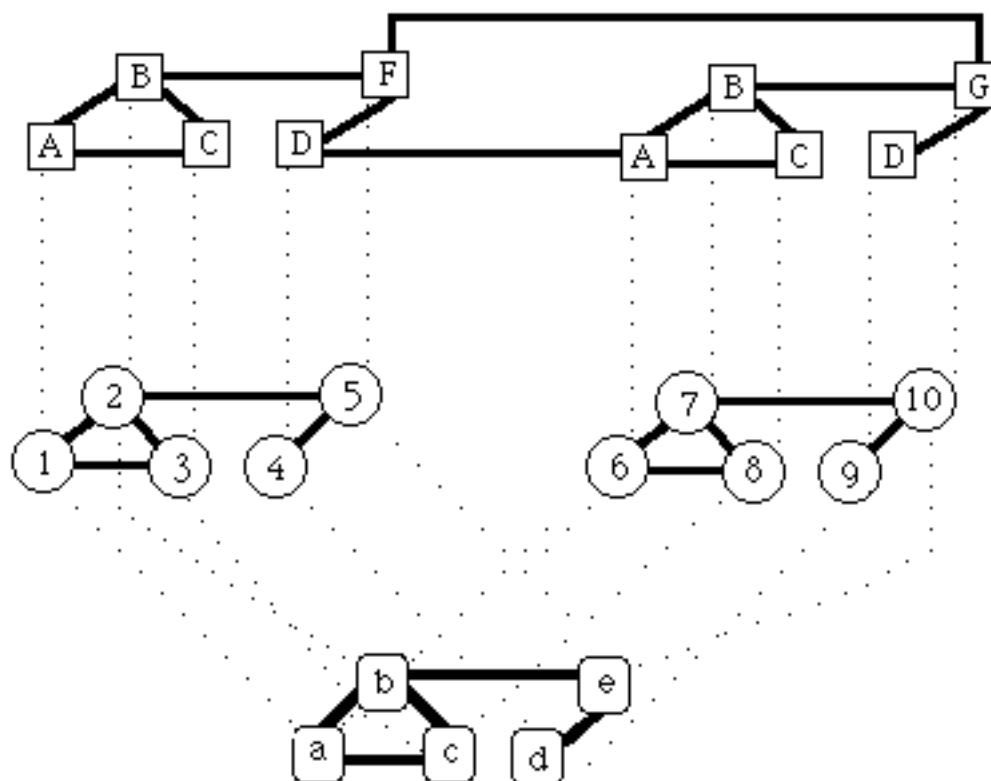


Figure 7.5: Interpretation of Twin Earth example.

The structure [(A) (B) (C) (D) (F)] represents Earth, while [(A) (B) (C) (D) (G)] represents Twin Earth, (F) and (G) representing water and XYZ, respectively. The relations between these upper structures is simply meant to convey the fact that both Earth and Twin Earth are parts of the same universe, and thus may bear non-interpretive relations to each other. Considered individually Earth and Twin Earth are similarly structured.

(1 - 5) and (6 - 10) represent the cognitive structures of me and my Twin, respectively. The states, of course, are numerically distinct, but the structures are isomorphic.

(a - e) represents 'my world', or the way I think the world is, which is exactly the same as how my Twin thinks the world is (thus its single representation in the diagram).

Notice that (5) and (10) both have content (e), but have different referents, (F) and (G) respectively.

Let's recast the Twin Earth objection in terms more directly applicable to GIST (see also Figure 7.5). Recall that I and my Twin have identical cognitive economies, and yet I

have water thoughts and he has XYZ thoughts. It seems as though nothing so far built into GIST can answer the objection. We might all agree that as far as *content* goes my Twin and I have similar thoughts. The worlds created by my world-emulator and his will be identical, and insofar as we are interpreting contents in terms of our individual worlds, there will be no difference. So far so good. But GIST cannot yet explain why such thoughts have a different interpretive *reference*. One might think that since the real world (i.e. the scenario set up in the thought experiment) contains both Earth and Twin Earth, it provides the means for distinguishing my thoughts from my Twin's. But this won't work, for why should I interpret water as the reference of *my* cognitive states and XYZ as the reference of my Twin's, and not vice versa? (That is, Why do the interpretation lines in Figure 7.5 go vertically from the middle to the top structures, rather than cross?) Since the relevant structural isomorphisms are the same, I ought to be equally licensed to interpret my Twin's thoughts as water thoughts or as XYZ thoughts. There are two answers to this.

First, I think it plausible to assume that our talk of meaning and reference evolved to fill two related and often conflated needs, the need to know what (de re) so-and-so is talking about (reference), and the need to know what (de dicto) so-and-so thinks (content). If I'm right about this, then the Twin example might be addressed as follows. The thought experiment begins by setting up, in your mind, a world that contains two different planets, which contain two different substances. My thoughts are about water (de re) because you interpret them to be about water, and not XYZ. You interpret the situation this way because it would be natural for you to expect that my utterances give you information about water, and not about XYZ. When I tell you "I just conducted some unprecedented experiments and found that water shifts from ice-8 to ice-9 under conditions C," you do not, on the basis of that information alone, learn anything about XYZ. Under normal interpretive conditions, my utterances provide you with information about one extended entity in your world, and not another. Now of course, one natural way we determine what (de re) other

people's utterances will give you information about is via causal interactions. And to this degree I think that causal theories of reference, etc., are on to something. But what they are on to is not some 'true essence' of reference, but merely a reliable heuristic (among others) used constantly in our interpretive practices.¹⁰⁷ This heuristic is culturally endorsed because it works most of the time.

Another way of putting this point is that *reference is entirely a manifestation of interpretation*. I have a world, populated with entities and relations. I want to learn more about this world, and one source of information is your utterances. But before your utterances tell me anything about the inhabitants of my world, I must interpret them. When you utter "the psychic on the corner went out of business," I learn that the charlatan on the corner went out of business. The important point to see is that the reason you can refer to the charlatan by uttering "the psychic" is that I interpret your psychic-thoughts in such a way that they most directly affect my charlatan-thoughts. The next question is: Why do I interpret your words/thoughts this way rather than some other way? And the answer to this question may involve many factors that under normal circumstances conspire, but which can be induced, by clever philosophers, to work at cross-purposes. Causal relations, conversational implicatures, evolved functions, pragmatic considerations, shared concerns, and perhaps even certain innate capacities are some of the considerations at play here.¹⁰⁸

I think it might be easy to miss the point I am trying to make here, so at risk of tedium I will try to restate it slightly differently. There has been a great deal of intelligence and industry devoted in the literature to determining the nature of reference: causal,

¹⁰⁷ In fact, the point can be made more strongly than I am making it here, because as Putnam has pointed out, the question of what the causal relations are is itself a matter of interpretation.

¹⁰⁸ My reading of Wittgenstein (1958) is that he recognized that some such innate or genetical-developmental considerations must come into play, and that the point where such considerations come into play is exactly the point at which 'explanations come to an end.' In saying that explanations come to an end, he meant that rational explanations come to an end, i.e. that the process of justifying certain practices on the basis of other such practices cannot go on forever. Eventually we get to questions like "Why do you match *this* color sample to *this* apple?" Presumably the ability to see colors, and to group colors in certain ways, is an innate capacity which underwrites our ability to become socialized into various practices and normatively governed activities, but which cannot itself be 'justified' or explained by any such practices.

descriptive, etc. Such accounts are typically advertised as providing conditions for a phrase or thought to refer to some entity, such as "P refers to x iff C." I do not want to argue here that any of these theories is right or wrong. What I do want to argue is that we should reinterpret their point. These theories, I think, expose considerations we bring to bear in our interpretive practice of reference fixing, but **THE REFERENCE FIXING IS DONE BY US, NOT BY THE CONDITIONS**. Thus such theories tell us about *us*, our intuitions and practices, directly, and only tell us about reference *indirectly*.

So, schematically, we might replace the previous formula with these two:

P refers to x iff we interpret P as referring to X.

We interpret P as referring to x on the basis of conditions C.

This is, of course, just a way of being an anti-realist about reference. The question is: Does P refer to X because we interpret it that way on the basis of conditions C, or do we interpret it that way because, on the basis of conditions C, P refers to X? I mean this to be fully analogous to the standard realism debate: Is P true because we all believe it, or do we all believe it because it's true?

That was the first answer. The second is perhaps a bit closer to the mark. To determine reference, we ask ourselves: What would the interpretee say the reference of an expression was *if he knew what I know*? That is, we imagine what it would be like to be the interpretee, and then augment the interpretee's epistemic position with additional knowledge of some sort, and imagine how the interpretee would judge the situation. Or, to make the tie with ETM as explicit as possible, we emulate the interpretee. We imagine Twin Me as in the example, but who learns about Earth and water, and is then asked "Have you been referring to water or XYZ with your prior 'water' thoughts?" And presumably, Twin Me would say (and I think this because if I were Twin Me, I would say) something like "I was thinking about XYZ, not water."

Now is a good time to take stock of where we are. Brains, through processes of emulation, build worlds. Smith's world-emulator provides Smith with *his world*, as Jones' world-emulator provides Jones with *his world*. These world emulators, by themselves, are basically the lower span of the Tower Bridge. According to IS, we map the internal dynamics of these emulators onto the structure of the target domain (the real world) in order to assign meanings. To the degree such mappings fail, Smith and Jones fail to represent.

The first step to GIST involved acknowledging that Smith's and Jones' world emulators might diverge significantly from each other, and from the real world, without impugning their representational status. We just needed to be more liberal with what we let count as a target domain, allowing a variety of worlds to count as legitimate. **Content** is thus the evaluation of a state's semantics according to the world created by the emulator of which it is a part.

The second move from IS to GIST involves recognizing two different evaluative projects to which we may subject candidate representational entities. The first, **content attribution**, is just the sort of enterprise discussed above. The second, **referential attribution**, is where representational states are assigned to entities from some exogenous target domain, typically the 'real world' or the interpreter's world. To the degree that the worlds of the interpreter and interpretee match, this project will be straightforward, the limiting case being the semantic evaluations of one's own thoughts, where it seems content and reference must coincide, since the two interpretive projects lose their distinguishing features.

There are some benefits to looking at matters this way. First, I think that GIST, as articulated so far, helps explain many of the intuitions about interpretation. Specifically, I think that to the degrees that they are correct, interpretive principles such as rationality, charity, humanity, etc., fall out as special cases subsumed by GIST interpretation. Second, GIST allows us to see clearly, I think, why there are conflicting intuitions about the

representational status of, e.g., thermometers. As mentioned earlier, GIST is primarily an interpretive account of semantics, and it recognizes (so far) two separate constraints on interpretation. First, there are constraints of isomorphism that govern interpretive content, and second there are constraints that govern interpretive reference. In the case of a thermometer, the content-assigning project cannot get started -- there seems to be no clear way to construct a thermometer's world, and thus we can assign its states no content. However, the second project is quite easy -- I can interpret the thermometer's states to give me information about entities in *my* world, namely temperatures of objects or areas. Thermometer states, according to GIST, have no content, but they can refer. The complementary case might be a member of an exotic or primitive culture. Here the interpretee is easily interpreted as having a world for itself, but many entities within that world resist translation. This can be diagnosed as a case where the associated state has a content (*intra*-world ascription works) but fails to have a referent (*real*-world ascription fails). Homer's gods may fall into this category.

7.6.3 Constraints on interpretation (part two): Burge

As has been pointed out by many others, my view of the world is to a tremendous degree a product of others, especially those in my culture. In fact, I think we can distinguish at least three different senses in which my world view is socially conditioned. First, and least interestingly, much of what I know I have learned from others, and have not verified for myself. Examples of such items are easily produced, especially in the realm of historical facts. I said that this is the least interesting sense because this sort of dependency is not much different from the fact that much of my world view comes from my eyes. Other people, like microscopes, thermometers and tree rings, can provide me with information that I could not get otherwise. A second way in which my world view is socially conditioned, a little more interesting than the first, is that my society provides the

infrastructure that underpins almost any investigation I might undertake. Someone else makes the microscopes, the chemicals, the buildings, the computers, and so on. Both of these sorts of dependency are somewhat uninteresting and fail to uphold some of the more grandiose claims that proponents of social theories of knowledge propose.

There is a third, more interesting, sort of social dependence, which is at the heart of the Burge examples. There is a sense in which the very concepts and groupings I use to structure my world are the property, ultimately, of the society at large. My society has the right to correct or alter my usage of almost any concept I employ. If I claim that glass is a solid or that whales are fish, I run the risk of being intellectually buffeted until I conform to normal usage. Most adults have been sufficiently buffeted already that most of their usage is in rough agreement, in much the same way that a person in a large crowd will have little choice but to head in the same direction as her immediate neighbors.

I want merely to point out the manner, which is old news, in which our usage owes allegiance to the mob. This distributed nature of correct application is a sort of boon - it allows me to refer to elms and tachyons though I have a limited grasp on them. My tachyon representation, which resides in my world emulator, is relatively unarticulated. But part of the subtext surrounding it is that there are people who do have well-articulated tachyon representations, and **what allows *me* to say that mine is a tachyon representation is that I agree to give those experts power of attorney over it.** It is the fact that I implicitly agree to defer to the experts in case of conflict that I can be said to refer to tachyons. One could accurately call this willingness to defer one's 'reference ticket', in that someone who holds the same sorts of vague and incomplete beliefs about tachyons as I do, but who claims that the physicists are wrong in the case of conflict, would not, I think, be interpreted as referring to tachyons.

The point I want to make is that my world includes a lot of pointers to experts and normal usage which I may be unable to articulate any more than simply acknowledging

their existence. *And when it comes to others trying to understand my world for purposes of interpretation or assigning content, the interpreter is obliged to honor those same pointers.*

The reason why the first subject in Burge's example has false beliefs about arthritis (as opposed to true beliefs about arthritis*) is that he has pointers to experts and normal usage in *his* society, and his usage is in conflict with *those* norms. The thought experiment is set up so that we will see the infraction even when the subject does not.¹⁰⁹ When I imagine the world of the subject in these examples, that world includes the same sorts of expert-pointers that my own world has. When I imagine his world, when I imagine what it would be like to be the subject, when I (to make the connection as bluntly as possible) *emulate* the subject, I take over those pointers, and in the case at hand, I know that the subject would, upon learning that his usage differed from that of his community, recognize that he had made a mistake, because **I** would recognize **my** mistake in such circumstances.

7.6.4 Opacity

Let us now look briefly at how GIST/ETM deals with opacity phenomena. The account outlined here is by-and-large compatible with the account developed in Fauconnier (1985). Consider the following sentences.

- (6) Gottlieb thinks that Hesperus is a star.
- (7) Hesperus is phosphorus.
- (8) Gottlieb thinks that Phosphorus is a star.

As is well known, it is quite possible for (6) and (7) to be true, and yet (8) false. The problem would seem to be that Gottlieb isn't aware of the identity expressed in (7). This

¹⁰⁹ I think that legal/ethical terminology is entirely appropriate here. The commuter driving 70 mph on I-5 is guilty even if she is not caught or even seen, while her counterpart on the German autobahn may be behaving completely legally.

highlights the fact that the embedded phrase 'Hesperus is a star' in (6) is an opaque attribution, in that one cannot substitute co-referring expressions without risk of changing the truth value of the matrix proposition.

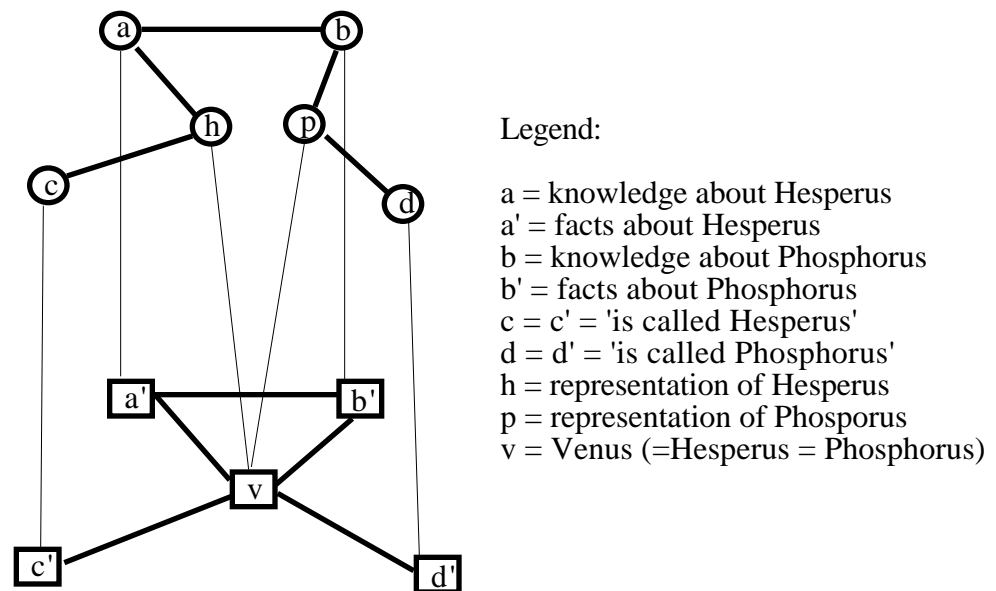


Figure 7.6: Imperfect interpretation and opacity.

Consider the diagram in Figure 7.6. Analogous to the previous illustrations, the circles at the top represent articulators in the 'world emulator' of Gottlieb. The thick lines connecting them represent causal (or cognitive) relationships obtaining between those articulators. The squares on the bottom represent 'real world' entities, including facts and social conventions that obtain (such as c', which is just the convention that Venus is sometimes called 'Hesperus'). The vertical lines are the interpretations. Notice that Gottlieb has separate representations for Hesperus and Phosphorus. The mysteriousness of opacity

vanishes as soon as we make the stunningly simple observation that reference does not determine content, and thus *co*-reference will not necessitate *co*-content.¹¹⁰

7.7 The Prospects for Communication

The solution to opacity phenomena in the last section appealed to differences between the world and the way people represent the world. It is thus to be expected that people will represent the world differently from each other. We have a solution to opacity, perhaps, but we have landed immediately back at the problem of the objectivity of content. If everyone gets his own little world within which content is characterized, then under what circumstances could we ever claim to be talking about the same thing, except in the exceedingly unlikely case where our respective worlds are identical?

We must recognize that part of what learning a language and becoming a member of a community involves is learning how to be in rough agreement with others over a large range of judgments. When mother told you that the cat was on the mat, if you had some strange notion that cats were all black and the thing she was pointing at was white, then you probably adjusted your world a bit to bring it more into accord with the apparent facts. Simple, ubiquitous workaday happenings like this force a degree of rough isomorphism between the 'worlds' of the members of a linguistic community. Thus in figure 7.6 above, though the lower and upper spans are not exactly isomorphic, there is a rough isomorphism, enough to allow Gottlieb's Hesperus and Phosphorus thoughts to be interpreted as referring to Venus. In some context, of course, where some of the outlying links are quite salient, we may not want to make that interpretation, because the portions of the structure under scrutiny may not be isomorphic enough for such an interpretation to be

¹¹⁰ Fauconnier (1985) develops a theory of tremendous elegance which deals not only with opacity phenomena, but with modals and counterfactuals, and others as well. I take it that the GIST account I have outlined here is entirely compatible with Fauconnier's account.

felicitous. The fact of the matter seems to be that communication and interpretation are best in exactly those 'simple' cases like 'The cat is on the mat,' and is most dicey in cases where such cultural assimilation is most vague (like the nature of justice, or beauty).¹¹¹

7.8 The role of the interpreter

A central concern for most theories of content is to provide a naturalistic account of why physical things can have an intrinsic meaning. The account I have provided may not seem to fill these criteria, since most of the puzzles of content were solved by appeal to an interpreter. And under the plausible assumption that interpretation is itself a semantic phenomenon, I have explained semantics in terms of semantics. Not a good naturalistic reduction. Or so it seems. There are a number of responses to this.

One response, which seems to be to be entirely sensible, and as far as I know, unappreciated, is that almost all of these 'problems of content' depend on problems of interpretation. In Putnam's Twin Earth example, he starts by explaining to the reader about two worlds, their similarities and differences. After the appropriate stage setting, the reader is then invited to admit that these two isomorphic structures (i.e. the word 'water' as uttered by me and my twin) have different meanings. But why should this be thought of as a problem of meaning at all, as opposed to a problem in interpretation? Notice that the problem only arises for the interpreter, in this case the reader. So what can be in principle wrong about providing an account, not of 'intrinsic content', but of interpretation, that explains why the reader is inclined to interpret Rick's utterances one way and Twin Rick's utterances some other way? To object that I have merely explained interpretation, and left

¹¹¹Similar points are made by P.M. Churchland (1979).

meaning unexplained, is simply to beg the question against the view that meaning *is* interpretation.

I'm not out of the woods yet, though. For now the objection continues: Fine, but what about the case where there are no interpreters around? On your (Grush's) view, meaning can only arise when there is some interpreter other than the agent around to interpret the agent's cognitive states. But this can't be right, for surely the agent's states are meaningful even if no one else is around. And furthermore, the objector continues, you still haven't finished off the 'alternate assignments' objection.

Fair enough. The reader may be pleased to know that I intend to close this chapter by answering these two final objections.

The first objection is, I think, easy to answer. It is that even if one is dealing with just a single agent, interpretation is in fact taking place. An agent is not just an emulator, emulators don't just exist on their own. Agents will typically have controllers that use emulators as stand-ins for the target. This is itself a sort of interpretation -- the controller, which is not itself representational (on my account) *interprets* the emulator *as* the target. And it can do this without itself representing the target because of the control structure (variability of the control loop) available. This could be called a sort of minimal interpretation. Furthermore (as was discussed briefly in Chapter Five), agents will typically have a number of models up and running, and will be interpreting some of them against others, for example to make sense of false belief. This, then, is my short answer to the genesis of meaning via natural interpretation: The use of emulators as control loop stand-ins provides a perfectly naturalistic account of interpretation, of what it means for one physical system to interpret another physical system as being about something other than itself.

Now to alternate assignments. My answer to this is that I really don't think it's a problem at all. Suppose that I was subjected to intense Brain-o-Scope scrutiny, of a sort that allowed a clever cognitive neuroscientist to map out the causal interactions among all

my relevant brain states. Just for fun, the experimenter comes up with some alternate assignments for these brain states: she determines, for example, that there is an alternate assignment according to which my 'cat' representation gets mapped onto ENVY instead of CAT. Though the scientist might think this is cute, there is no need for me to worry that the contents of my thoughts are in danger of evaporating. In order to convince me that this alternate assignment has been carried out right, I will need to be shown the entire alternate assignment scheme. And there is no reason for me to take this alternate assignment scheme as any more than a dictionary for some jury-rigged language in which the orthographic term 'ENVY' means 'cat.'

This is a subtle but powerful point that merits repetition. In order to make an alternate assignment, the scientist must find or create some putative structure which will be appropriately isomorphic to my cognitive economy. But what makes my cognitive economy *cognitive* is the fact that it supports world-like structure. Hence this new isomorphism will also be a world-like structure. So, when informed that my 'cat' tokens play a structurally isomorphic role to 'envy' tokens under this new mapping supported by this new structure, the most natural interpretation will have to be that 'envy' means 'cat.' And similarly for all such possible assignments.

7.9 Conclusion

I said earlier in this chapter that the neo-Fregeans and the x-role theorists were each right and wrong about some aspects of semantics. And furthermore, I said that the final version I would arrive at would treat meaning as (more or less) empirically neo-Fregean, and as transcendently x-role theoretic. I want to unpack this in the following way. By 'empirically neo-Fregean,' I mean that given a world populated by objects, relations, properties and the like, and given also the interpretive practices of cognitive agents like

ourselves, what the meaning of a cognitive token or linguistic item is, is just the thing or relation or property that it is interpreted as standing for. The interpretation diagrams used in this chapter exemplify this. To this extent the neo-Fregeans are right.

By 'transcendentally x-role theoretic' I mean something that needs a bit more explanation. Neo-Fregean semantics requires, as I just mentioned, that a world of some sort be given. It is this world which provides the ontology, and hence the extensions, that will be the meanings of the semantically evaluable tokens under consideration. But when we ask the further question: Where does this (or do these) worlds come from? The neo-Fregeans have no answer, aside from adoption of some sort of realism. What I want to say, indeed what I have argued for, is that these worlds are the product of the structures in which cognitive tokens, or emulator articulators, play a role. These complex cognitive economies provide world-like structure, by being organized in certain ways, and it is this world-like structure that is the precondition for the possibility of content.

References:

- Arbib, M. (1981) 'Perceptual structures and distributed motor control' in Brooks, V.B., ed. Handbook of Physiology: The Nervous System, II. Motor Control. American Physiological Society. Bethesda, pp. 1448-1480.
- Astington, J. and Gopnik, A. 'Knowing you've changed your mind: Children's understanding of representational change' in Astington, J., Harris, P., and Oleson, D. eds. (1988) Developing theories of mind New York: Cambridge University Press
- Astington, J., Harris, P., and Oleson, D. eds. (1988) Developing theories of mind New York: Cambridge University Press
- Atekeson, C.G. and Reinkensmeyer, D.J. (1990) 'Using associative content-addressable memories to control robots' in Miller, W.T., Sutton, R.S. and Werbos, P.J., eds. Neural Networks for Control MIT Press/Bradford
- Block, Ned (1987) "Advertisement for a semantics for psychology" in French, Uehling and Wettstein, eds. Midwest Studies in Philosophy X University of Minnesota Press
- Brooks, R. (1991) 'Intelligence without representation' Artificial Intelligence 47:139-159
- Burge, Tyler (1979) 'Individualism and the Mental' Midwest Studies in Philosophy 4:73-121
- Chomsky, Noam (1981) Lectures on Government and Binding Dordrecht, Foris.
- Chomsky, Noam (1986) Barriers MIT/Bradford
- Churchland, Paul (1979) Scientific Realism and the Plasticity of Mind Cambridge: CUP
- Churchland, Paul (1989) A neurocomputational perspective MIT Press/Bradford
- Clark, Andy (1994) Associative Engines MIT Press/Bradford
- Cummins, R. (1989) Meaning and mental representation MIT Press/Bradford
- Davidson, Donald (1973) 'Radical Interpretation' in Inquiries into Truth and Interpretation (1984) Clarendon Press, Oxford

- Deane, Paul (1988) 'Which NPs are there unusual possibilities for extraction from?' in MacLeod, L., Larson, G., and Brentari, D. eds. CLS 24: Papers from the 24th Annual Regional Meeting of the Chicago Linguistics Society -- Part One: The General Session. Chicago: Chicago Linguistics Society
- Deane, Paul (1991) 'Limits to attention: A cognitive theory of island constraints' Cognitive Linguistics 2(1):1-63
- Deane, Paul (1992) Grammar in the Mind and Brain: Explorations in Cognitive Syntax Mouton de Gruyter
- Decety, J. and Michel, F. (1989) 'Comparative analysis of actual and mental movement times in two graphic tasks' Brain and Cognition 11:87-97
- Decety, J., Sjöholm, H., Ryding, E., Stenberg, G., and Ingvar, D. (1990) 'The Cerebellum participates in Cognitive Activity: Tomographic measurements of regional cerebral blood flow' Brain Research 535: 313-317
- Decety, J., Jeannerod, M., Germain, M. and Pastene, J. (1991) 'Vegetative response during imagined movement is proportional to imagined effort' Behavioral Brain Research 42:1-5,
- Denier van der Gon (1988) "Motor control: Aspects of its organization, control signals and properties" in Wallinga et al. eds. Proceedings of the 7th Congress of the International Electrophysiological Society
- Dennett, Daniel (1987) The intentional stance MIT Press/Bradford
- Dretske, Fred (1981) Knowledge and the flow of information MIT Press/Bradford
- Dretske, Fred (1983) 'Precis of *Knowledge and the Flow of Information*' Behavioral and Brain Sciences 6:55-90
- Farah, M., Sos, M., and Dasheiff, R. (1992) 'Visual angle of the mind's eye before and after unilateral occipital lobectomy' Journal of Experimental Psychology: Human Perception and Performance 18:241-246
- Fauconnier, Gilles (1985) Mental spaces MIT Press/Bradford
- Feltz, D.L. and Landers D.M. (1983) 'The effects of mental practice on motor skill learning and performance: a meta-analysis' Journal of Sport Psychology 5:25-57
- Fodor, J. (1975) The language of thought Harvard University Press
- Fodor, J. (1987) Psychosemantics MIT Press/Bradford
- Fodor, J. (1990) A theory of content MIT Press/Bradford
- Fodor, J. and Pylyshyn, Z. (1988) Connectionism and cognitive architecture: a critical analysis Cognition 28 (1-2) 3-71

- Fox, P.T., Pardo, J.V., Petersen, J.V. and Raichle, M.E. (1987) 'Supplementary motor and premotor responses to actual and imagined hand movements with positron emission tomography' Neuroscience Abstracts 1433
- Frege, G. (1952) 'On sense and reference' in Geach, P. and Black, M. eds. Translations from the Philosophical Writings of Gottlob Frege Oxford: Blackwell
- Ghez, C. (1990) 'Voluntary Movement' in Kandel, E., Schwartz, J., and Jessell, T. eds. Principles of Neural Science, Third Edition. Elsevier
- Ghez, C. and Vicaro, D. (1978) 'Discharge of red nucleus neurons during voluntary muscle contraction: activity patterns and correlations with isometric force' Journal of Physiology, Paris 74:283-285
- Goldman, A. (1993) 'The psychology of folk psychology' Behavioral and Brain Sciences 16:15-28
- Goodwin, G.M., McCloskey, D.I. and Mitchell, J.H. (1972) 'Cardiovascular and respiratory responses to changes in central command during isometric exercise at constant muscle tension' Journal of Physiology 226: 173-190
- Gordon, Robert (1986) 'Folk psychology as simulation' Mind and Language 1:158-71
- Gray, C., and Singer, W. (1989) 'Stimulus-specific neural oscillations in orientation specific columns of the visual cortex' Proceedings of the National Academy of Science 86:1698-1702
- Grush, R. (1994a) 'Motor models as steps to higher cognition' Behavioral and Brain Sciences 17:2:209-210, commentary on Jeannerod (1994)
- Grush, R. (1994b) 'Beyond connectionist vs. classical AI: A control theoretic perspective on development and cognitive science' Behavioral and Brain Sciences 17:4:720 commentary on Karmiloff-Smith, A. (1994) Precis of Beyond Modularity: A developmental perspective on cognitive science Behavioral and Brain Sciences 17:4:
- Haegeman, Liliane (1991) Introduction to government and binding theory Blackwell
- Held, R. and Hein, A. (1963) 'Movement-produced stimulation in the development of visually guided behavior' Journal of Comparative and Physiological Psychology 56:5:872-876
- Houk, J.C. (1988) 'Schema for motor control using a network model of the cerebellum' in Anderson, D.Z. ed. Neural Information Processing Systems 367-376 New York: American Institute of Physics
- Houk, J.C., Singh, S.P., Fischer, C., and Barto, A. (1990) 'An adaptive sensorimotor network inspired by the anatomy and physiology of the cerebellum' in Miller, W.T., Sutton, R.S. and Werbos, P.J., eds. Neural Networks for Control MIT Press/Bradford

- Ingvar, D. and Philipsson, L. (1977) 'Distribution of the cerebral blood flow in the dominant hemisphere during motor ideation and motor performance' Annals of Neurology 2:230-237
- Ito, Masao (1984) The Cerebellum and Neural Control Raven Press
- Jackendoff, R. (1990) Semantic Structures MIT Press/Bradford
- Jeannerod, M. (1994) 'The representing brain - Neural correlates of motor intention and imagery' Behavioral and Brain Sciences 17:2:187-202
- Johnson, Mark (1987) The body in the mind University of Chicago Press
- Johnson-Laird, P. (1983) Mental Models Harvard University Press
- Karmiloff-Smith, A. (1992) Beyond Modularity: A developmental perspective on Cognitive Science MIT Press/Bradford
- Kawato, M. (1989) 'Adaptation and learning in control of voluntary movement by the central nervous system' Advanced Robotics 3:3:229-249
- Kawato, M. (1990) 'Computational schemes and neural network models for formation and control of multijoint arm trajectories' in Miller, W.T., Sutton, R.S. and Werbos, P.J., eds. Neural Networks for Control MIT Press/Bradford
- Kawato, M. Furukawa, K. and Suzuki, R. (1987) 'A hierarchical neural network model for control and learning of voluntary movement' Biological Cybernetics 57:447-454
- Kluender, R. (1990) 'A neurophysiological investigation of wh-islands' in Hall, K., Koenig, J., Meacham, M., Reinma, S., and Sutton, L., eds. Proceedings of the 16th Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on the Legacy of Grice. Berkeley: Berkeley Linguistics Society
- Lakoff, G. (1986) 'Frame semantic control of the coordinate structure constraint' in Farley, A., Farley, P. and McCulloch, K. eds. CLS 22 Part 2: Papers from the parasession on pragmatics and grammatical theory. Chicago: Chicago Linguistics Society.
- Lakoff, George (1987) Women, fire and dangerous things University of Chicago Press
- Langacker, Ronald (1987) Foundations of Cognitive Grammar: Volume One Stanford University Press
- Langacker, Ronald (1990) Concept, Image and Symbol Mouton de Gruyter
- Langacker, Ronald (1991) Foundations of Cognitive Grammar: Volume Two Stanford University Press
- Leslie, Alan (1988) 'Some implications of pretense for mechanisms underlying the child's theory of mind' in Astington, J., Harris, P., and Oleson, D. eds. Developing theories of mind New York: Cambridge University Press

- Llinas, R., and Pare, D. (1991) "On dreaming and wakefulness" Neuroscience 44:3:521-535
- Mandelblit, N. (1993). "Machine Translation: a Cognitive Linguistics Approach". In Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan, 1993.
- McClelland, J., Rumelhart, D. and the PDP Research Group (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume 2 MIT Press/Bradford
- Mel, Bartlett (1986) "A connectionist model for learning 3-d mental rotation, zoom and pan" In Proceedings of the 8th Annual Conference of the Cognitive Science Society
- Miller, W.T., Sutton, R.S. and Werbos, P.J., eds.(1990) Neural Networks for Control MIT Press/Bradford
- Millikan, Ruth (1984) Language, thought, and other biological categories MIT Press/Bradford
- Nguyen, D. and Widrow, B. (1989) 'The truck backer-upper: An example of self-learning in neural networks' in Proceedings of the International Joint Conference on Neural Networks II 357-363 IEEE Press, New York
- Perner, J. (1988) 'Developing semantics for theories of mind: From propositional attitudes to mental representations in Astington, J., Harris, P., and Oleson, D. eds. Developing theories of mind New York: Cambridge University Press
- Perner, J., Leekham, S.R., and Wimmer, H. (1987) 'Three-year-olds' difficulty with false belief' British Journal of Developmental Psychology 5:125-137
- Piaget, J. (1962) Play, dreams and imitation in childhood W.W. Norton and Company
- Piaget, J. and Inhelder, B. (1969) The Psychology of the Child Basic Books
- Putnam, H. (1975) The meaning of 'meaning' in Mind, Language and Reality: Philosophical Papers Volume 2 Cambridge University Press
- Putnam, H. (1983) Realism and Reason: Philosophical Papers Volume 3 Cambridge University Press
- Pylyshyn, Z. (1984) Computation and Cognition MIT Press/Bradford
- Requin, J., Brenner, J. and Ring, C. (1991) 'Preparation for Action' in J.R. Jennings and M. Coles, eds. Handbook of Cognitive Psychophysiology: Central and Autonomic nervous system approaches John Wiley and Sons
- Ross, J.R. (1987) 'Islands and syntactic prototypes: in Need, B., Schiler, E., and Bosch, A. eds. Papers from the general session at the 23rd annual regional meeting of the Chicago Linguistic Society CLS 23 part 1.309-20
- Rizzi, Luigi (1990) Relativized minimality MIT Press/Bradford

- Roland, P.E., Larsen, B., Lassen, N.A., and Skinhoj, E. (1980) 'Supplementary Motor area and other cortical areas in organization of voluntary movements in man' Journal of Neurophysiology 43: 1:118-136
- Rumelhart, D. and McClelland, J. and the PDP Research Group (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume 1 MIT Press/Bradford
- Searle, John (1980) "Minds, Brains and Programs" Behavioral and Brain Sciences 3:417-24
- Shastri, L., and Ajjanagadde, V. (1993) 'From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony' Behavioral and Brain Sciences 16:417-494
- Singer, W. (to appear) 'Neuronal synchronization: A solution to the binding problem?'
- Stuurman, F. (1985) Phrase structure theory in generative grammar Dordrecht: Foris
- Tsukahara, N. and Kawato, M. (1982) 'Dynamic and plastic properties of the brain stem neuronal networks as the possible neuronal basis of learning and memory' in S. Amari and M. Arbib eds. Lecture notes in biomathematics Volume 45: Cooperation and competition in neural nets, 430-441 New York, Springer-Verlag
- Unger, L. (1990) 'A bioreactor benchmark for adaptive network-based process control' in Miller, Sutton and Werbos, eds. (1990) Neural Networks for Control MIT Press/Bradford
- van der Meulen, J.H.P, Gooskens, R.H.J.M., Dennier van der Gon, J.J., Gielen, C.C.A.M., and Wilhelm, K. (1990) 'Mechanisms underlying accuracy in fast goal-directed arm movements in man' Journal of Motor Behavior 22:1:67-84
- van Gelder, T (1991) 'Connectionism and Dynamical explanation' Proceedings of the 13th Annual conference of the Cognitive Science Society Hillsdale, Earlbaum pp. 499-503
- van Hoek, Karen (1992) Paths through conceptual structure: Constraints on pronominal anaphora [unpublished PhD dissertation, University of California, San Diego]
- von der Malsburg, C., and Buhmann, J. (1992) 'Sensory segmentation with coupled neural oscillators' Biological Cybernetics 67:233-242
- von der Malsburg, C. and Schneider, W. (1986) 'A neural cocktail party processor' Biological Cybernetics 54:29-40
- Wang, Y. and Morgan, W.P. (1992) 'The effects of imagery perspectives on the physiological responses to imagined exercise' Behavioral and Brain Research 52:167-174

- Wellman, H. (1988) 'First steps in the child's theorizing about the mind' in Astington, J., Harris, P., and Oleson, D. eds. Developing theories of mind New York: Cambridge University Press
- Wellman, H. (1990) The child's theory of mind MIT Press/Bradford
- Wimmer, H. and Perner, J. (1983) 'Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception' Cognition 13:103-128
- Wittgenstein, L. (1958) Philosophical Investigations Macmillian Publishing
- Yue, G. and Cole, K.J. (1992) 'Strength increases from the motor program. Comparison of training with maximal voluntary and imagined muscle contractions' Journal of Neurophysiology 67:1114-1123