The Integration of Learning into Models of Human Memory

Simon John Dennis BSc (Hons), BA

A thesis submitted for the degree of Doctor of Philosophy

Department of Computer Science

The University of Queensland

March 28, 1994

MEMORY! Always there, of course, but usually hidden. And then, sometimes, as a result of just the right kind of push, it could emerge suddenly, sharply defined, all in color, bright and moving and alive.

Robots and Empire, Issac Asimov

Statement of Originality

The work presented in this thesis is, to the best of my knowledge and belief, original, except as acknowledged in the text, and the material has not been submitted, either in whole or in part, for a degree at this or any other university.

Simon Dennis,

March 28, 1994

Abstract

Since memory was first distinguished as a separate phenomenon from learning (Melton, 1963), researchers in the area have concentrated on the memory component. Mathematical models, such as SAM (Raaijmakers & Shiffrin, 1981; Gillund & Shiffrin, 1984), TODAM (Murdock, 1982), CHARM (Eich, 1982), Minerva II (Hintzman, 1984) and the matrix model (Pike, 1984; Humphreys, Bain, & Pike, 1989), have focussed on the mechanisms of encoding, storage and retrieval. The affects of variables such as retention time, number of presentations, spacing of presentations, type of retrieval test, nature of cues, encoding paradigm and the extent to which the study context is specified in the test instructions have been studied empirically and modelled.

The learning component, which was the focus of the field for much of this century, has received less attention in recent years (Estes, 1991). Despite the extensive empirical database on learning phenomena (Postman, Burns, & Hasher, 1970), attempts to model this data have been few. The attempts that do exist have concentrated on specifying algorithms by which experience might tune the parameters of existing memory models (Murdock, 1987) rather than attempting to explain how learning induces the representations, decision criteria and control processes of memory in the first instance.

How people acquire these components of the memory system has important ramifications for the study of human retention. One of the most critical of these ramifications is the nature of the relationship between the environment and the mechanism of memory. Recent empirical work on the the environment of memory has revealed a striking correspondence between the structure of the environment and the pattern of performance in human subjects (Anderson & Schooler, 1991). This thesis extends this work by studying the environment empirically, developing a learning mechanism and demonstrating that this learning mechanism behaves in a qualitatively similar fashion to human subjects when exposed to an environment

that mirrors that with which the subjects contend.

Analyses of the relevant environments of two touchstone phenomena: the list strength effect in recognition and the word frequency effect in recognition were performed to establish the context in which interactive accounts of these phenomena must be set. It was found that while low frequency words occur less often than high frequency words, they are more likely to recur within a context. In addition, the probability of recurrence was found to increase if a word had occurred frequently in the current context, but was not affected by the amount of repetition of words other than the target word.

A learning or *interactive* model of human memory called the Hebbian Recurrent Network (HRN) has been developed. The HRN integrates work in the mathematical modelling of memory with that in error correcting connectionist networks by incorporating the matrix model (Pike, 1984; Humphreys et al., 1989) into the Simple Recurrent Network (SRN, Elman, 1989, 1990). The result is an architecture which has the desirable memory characteristics of the matrix model such as low interference and massive generalisation but which is able to *learn* appropriate encodings for items, decision criteria and the control functions of memory which have traditionally been chosen a priori in the mathematical memory literature. Simulations demonstrate that the HRN is well suited to the recognition task. When compared with the SRN, the HRN is able to learn longer lists, generalises from smaller training sets, and is not degraded significantly by increasing the vocabulary size.

To demonstrate that the HRN learning mechanism is capable of addressing experimental behaviour, the phenomena studied environmentally were modelled with the HRN. The HRN showed a low frequency word advantage when it was presented with an environment in which high frequency words occurred more often, but low frequency words were more likely to recur within a context. In addition, the HRN showed a null list strength effect while retaining the list length and item strength effects when exposed to an environment in which the environmental results were embedded.

By incorporating a learning mechanism and examining the environment in which memory models are situated it is possible to produce models that: (1) can start to address developmental phenomena; (2) can provide a mechanism to address learning-

to-learn phenomena; (3) can address how internal states attain their meanings; (4) are easily extended to a wide variety of cognitive phenomena; and (5) account for the striking similarity between the environmental demands placed upon the memory system and the performance of human subjects.

List of Publications

S Dennis. Adding memory to recurrent networks: The Hybrid Recurrent Network. In S Goss, editor, *Proceedings of the 2nd Australian Cognitive Science Conference*, pages 58–61, 1993.

S Dennis and J Wiles. Introducing learning into models of human memory: The Hebbian Recurrent Network. In *Program of the Fifteenth Annual Conference of the Cognitive Science Society*, pages 394–399. Lawrence Erlbaum Associates, 1993.

S Dennis and J Wiles. Introducing learning into models of human memory: The Hebbian Recurrent Network. Technical Report 252, Department of Computer Science, The University of Queensland, Australia, January 1993.

S Dennis, J Wiles, and M S Humphreys. What does the environment look like? Setting the scene for interactive models of human memory. Technical Report 249, Department of Computer Science, The University of Queensland, Australia, November 1992.

M S Humphreys, J Wiles, and S Dennis. Data structures and access processes: A first approximation to a theory of human memory. Submitted to Behavioral and Brain Sciences, 1992.

J Wiles, M S Humphreys, J D Bain, and S J Dennis. Direct memory access using two cues: Finding the intersection of sets in a connectionist model. In D S Touretsky and R Lippmann, editors, *Advances in Neural Information Processing Systems 3*, pages 635–641, San Mateo, CA, 1991. Morgan Kaufmann.

S Dennis, M S Humphreys, and J Wiles. Memory access using multiple independent cues: An additive model. In Marwan Jabri, editor, *Proceedings of the*

Second Australian Conference on Neural Networks, pages 49–54, 1991.

S Dennis and S Phillips. Analysis tools for neural networks. Technical Report 207, Department of Computer Science, The University of Queensland, Australia, May 1991. In conjunction with this technical report we produced three programs which implemented Principle Components Analysis, Canonical Discriminant Analysis, and Group Error Analysis, and developed a tutorial. This package, called NetTools is now distributed with the Aspirin/Migraines neural network simulator.

S Dennis. A multilevel reinforcement learning rule. In Marwan Jabri, editor, *Proceedings of the First Australian Conference on Neural Networks*, pages 99–102, 1990.

Acknowledgments

The completion of a thesis is an important event in the life of many people including the writer and I am pleased to be able to thank those who have contributed.

- My wife, Alison, who has been a source of encouragement and strength throughout.
- My family, for their moral support both during the PhD and through my formative years.
- My supervisor, Dr. Janet Wiles for her advice, encouragement and support during the PhD process. The presentation and readability of the thesis are in large part attributable to her critical eye.
- Professor Mike Humphreys, who has contributed greatly to the psychological grounding of the thesis.
- Professor Richard Shiffrin, with whom I spent an enjoyable and particularly creative winter (1991) at the University of Indiana being educated in the art of Tibetan dining and the list strength effect.
- Steven Phillips for many helpful comments and long discussions on systematicity and the implications of encodingism (see appendix A).
- The members of the Connectionist Research Group, Glenda Andrews, Paul Bakker, Mark Chapell, Kerry Chalmers, Julie Stewart, Hanna Majewski, Guy Smith, Kate Stevens, and Dr. Helen Purchase for a forum in which to present ideas and for the helpful discussion which ensued.
- The memory research group, Department of Psychology, University of Indiana, and particularly Peter Noble for many interesting discussions.

- Dr. Tony Baglioni for advise on the statistical analysis particularly in chapter two.

- Dr. Ellen Watson and Milos Novak for thoughtful discussion about the philosophical aspects of the thesis.

- The Commonwealth Department of Employment, Education and Training for financial support through an Australian Postgraduate Research Award.

- The Department of Computer Science, University of Queensland for resources, financial support and a position as tutorial assistant.

Contents

Statement of Originality ii

Abstract iii

List of Publications vi

Acknowledgments viii

Contents x

List of Figures xv

List of Tables xvii

Preface xviii

1 Introduction: Learning in Human Memory 1

1.1 The Status of Learning in Models of Human Memory 1

1.2 Learning Defined 5

1.3 Review of Current Models of Recognition Memory 7

1.3.1 Theory of Distributed Associative Memory 7

1.3.2 Search of Associative Memory 8

1.3.3 The Matrix Model 10

1.3.4 Minerva II 11

1.3.5 The Multilayer Backpropagation Network 12

1.4 Why learning is important in memory 14

1.4.1 Learning and Development 15

1.4.2 Learning-to-Learn 15

1.4.3 Learning and Meaning 17
1.4.4 Learning and Cross Paradigm Generality 17
1.4.5 Learning and the Environment 18
1.5 Role of Environment in Memory Models 18
1.5.1 Evolution and the Rational Analysis of Memory 20
1.5.2 Learning and the interactive approach 23
1.6 Outline of the Thesis 23

2 The Environment of Memory **27**
2.1 Word Frequency in Recognition 28
2.1.1 Experimental Results and Current Accounts 28
2.1.2 Environmental Analysis 33
2.2 The Null List Strength Effect in Recognition 42
2.2.1 Experimental Results and Current Accounts 42
2.2.2 Environmental Analysis 45
2.3 Towards an Environmental Analysis of Cued Recall 54
2.4 Discussion and Conclusions 56

3 The Mechanism of Memory **59**
3.1 Learning Issues 61
3.1.1 Learning Representation 61
3.1.2 Learning Decision Criteria 62
3.1.3 Learning Control 63
3.2 Memory Issues 63
3.2.1 Capacity and Interference 64
3.2.2 Generalisation 64
3.2.3 Rapid Binding 65
3.3 The Hebbian Recurrent Network 66
3.3.1 The Objectives 66
3.3.2 The Model 66
3.3.3 The Task: Episodic Recognition 74
3.3.4 Stepping through a Simple Example 74
3.4 Simulation Method 78

3.4.1 Measures 78
3.5 Results and Discussion 79
3.5.1 List Length 79
3.5.2 The Memory Decay Parameter 83
3.5.3 Generalisation 85
3.5.4 Vocabulary Size 86
3.5.5 Time Complexity 86
3.6 Hidden Unit Analysis 88
3.7 Conclusions 92

4 Word Frequency Example **99**
4.1 Introduction 99
4.2 The Environmental Context 101
4.3 Simulation 1: Word frequency and the HRN 101
4.3.1 Method 102
4.3.2 Results 107
4.3.3 Discussion 108
4.4 Simulation 2: Adding Weight Decay 110
4.4.1 Method 110
4.4.2 Results 110
4.4.3 Discussion 112
4.5 Conclusions 113

5 List Strength Example **115**
5.1 Introduction 115
5.2 The Environmental Context 116
5.3 Simulation: List strength and the HRN 118
5.3.1 Method 119
5.3.2 Results 122
5.4 Discussion and Conclusions 123

6 Conclusion and Further Work **127**
6.1 Summary of the Contribution 127
6.2 Evaluation and Further Work 129

6.2.1 Learning and Development 129
6.2.2 Learning-to-Learn 130
6.2.3 Learning and Meaning 132
6.2.4 Learning and Cross Paradigm Generality 133
6.2.5 Learning and the Environment 139

Bibliography **141**

A Learning and Meaning **159**

B Example of Environmental Data Processing **165**
B.1 Raw Text Sample 165
B.2 Clean Text Sample 168
B.3 The Segments Extracted from the Clean Text Sample 170

C Regression of List Strength Data **171**

D Rational Analysis of Selected Phenomena **173**
D.1 The Word Frequency Effect in Recognition 173
D.2 The List Strength Effect in Recognition 175
D.3 The ABABr Paradigm 177
D.4 The Crossed Associates Paradigm 180

E Verifying the List Strength Training Set **183**
E.1 Expressing the environmental constraints 183
E.2 The training set statistics 185
E.3 The item strength constraint 186
E.4 The length constraint 187
E.5 The null list strength constraint 188
E.6 Conditional of Ors Identity 189

Glossary **191**

Index **193**

List of Figures

1.1 The robot arm framework for understanding the relationship between adaptive processes. 19
1.2 Correspondence relationships between rational analysis and ACT*. . 22
1.3 Map of the thesis. 25

2.1 Operating characteristics for physics and arts students on physics and common words. 31
2.2 Word density versus word frequency. 38
2.3 Mean word density verses word frequency. 39
2.4 An idealised graph of the word frequency by occurrence history interaction. 40
2.5 Recurrence probability of a word verses time since its last occurrence. 41
2.6 Differentiation in the SAM model. 44
2.7 Recurrence probability as a function of frequency. 47
2.8 Recurrence probability as a function of recency. 48
2.9 The division of articles in the environmental study of list strength. . . 51

3.1 The Hebbian Recurrent Network (HRN). 70
3.2 Memory processes in the HRN. 72
3.3 The trained HRN processing a target and a distractor. 76
3.4 Performance (D') as a function of list length on the HRN and SRN. . 80
3.5 Learning curves for the SRN as list length was varied. 81
3.6 Learning curves for the HRN as list length was varied. 82
3.7 Orthogonality of hidden unit patterns of the HRN as a function of training time. 84

3.8 Performance (D') of the HRN as a function of the memory decay parameter on the 4-Item recognition task. 84
3.9 Performance (D') of the HRN as a function of item position for three values of the memory decay parameter. 85
3.10 Performance of the SRN and HRN as a function of number of training examples. 86
3.11 Performance (D') of the HRN as a function of the number of vocabulary items on the 4-item recognition task. 87
3.12 Time complexity of the HRN on the recognition task. 87
3.13 Principal components plot of the hidden patterns of the HRN. 89
3.14 Canonical discriminants plot of the hidden patterns of the HRN grouped on position. 90
3.15 Canonical discriminants plots of the hidden patterns of the HRN grouped on input and on position. 91
3.16 Canonical discriminant plots of the hidden unit patterns of the SRN and HRN grouped on output. 93
3.17 Hierarchical Cluster Analysis (HCA) of the hidden unit patterns of the SRN after the study list has been input. 94
3.18 Hierarchical Cluster Analysis (HCA) of the hidden unit patterns of the HRN after the study list has been input. 95

4.1 The HRN architecture for the word frequency simulations. 103
4.2 Training epoch verses performance (D') of the HRN on high and low frequency words. 108
4.3 Learning curves for the HRN applied to the recognition task for several values of the weight decay parameter. 111

5.1 The HRN architecture for the list strength simulations. 120
5.2 Training epoch verses performance (D') of the HRN for each of the list strength conditions. 124
5.3 Training epoch verses the list strength ratio of the HRN. 125

List of Tables

2.1 Accuracy results for nonrepeated items in different list types. 43

3.1 The HRN algorithm. 71

3.2 An example input/output sequence for the episodic recognition task. 75

4.1 Training set statistics for the word frequency simulation. 105

5.1 Example target lists from the list strength training set. 121

6.1 Example of lists in the ABAB' paradigm. 131

C.1 Significance levels of adding each of the independent variables on the prediction of recurrence probability for each of 20 replications. 171

E.1 Probabilities of each of the list types in the list strength training set. 185

E.2 Probabilities that a given word appears in a given condition in the list strength training set. 185

E.3 Probabilities that a word occurs as a test item given the condition in the list strength training set. 185

Preface

Just as the tennis player acquires a backhand or a golfer develops their putting, the skills of memory are learned. Children learn strategies to help them remember everyday events such as cricket matches or trips to the swimming pool. The chess master learns to recognise the structure of legal board positions and the student learns to organise the material they need to recall for their examinations. Memory is a collection of cognitive skills (Kail, 1984) and these skills are learned.

Existing computational models[1] provide detailed accounts of many aspects of human memory. They describe the affect on memory performance of variables such as item frequency, item lag and item spacing. They account for the affect of stimulus variables such as word frequency and number of associates. They capture important insights into the difference between processes such as recall and recognition. But they do not learn. The storage and retrieval processes are considered fixed and neither the developmental course of memory skills, nor the improvement of memory skills in adults is addressed.

Modelling the acquisition of memory processes requires the introduction of a mechanism that promotes a role for the environment. The environment is not only the source of inputs and the destination of the outputs, but plays a causal role in cognition. In "The Adaptive Character of Thought", John Anderson proposes that there is much to be gained by reflecting on the environment of the human cognitive system.

> The human is not a random construction but a construction that has been, to some degree, optimised to its environment by evolution. The behaviour computed by our cognitive mechanisms must be optimised to some degree and in some sense. If we could only specify that degree and

[1]Such as SAM (Raaijmakers & Shiffrin, 1981; Gillund & Shiffrin, 1984), TODAM (Murdock, 1982), CHARM (Eich, 1982), Minerva II (Hintzman, 1984) and the matrix model (Pike, 1984; Humphreys et al., 1989).

> sense, we would be in a position to place enormous constraints on our proposals for cognitive mechanisms, both at the implementation level and the algorithmic level."

(Anderson, 1990, p 26)

The mechanisms of cognition are not the result of blind chance, but are the culmination of a process of adaption to the stream of environmental stimuli. Given that the environment is more amenable to observation than the mechanisms, the statistical landscape of the environment is a potentially valuable source of constraint which has only just begun to be tapped by cognitive scientists (Anderson, 1990; Anderson & Schooler, 1991; Dennis, Wiles, & Humphreys, 1992). This thesis considers what is gained by applying these constraints in the context of human memory.

As Belew has said, "It is possible to view the interacting processes of evolution, learning, and culture as adaptive algorithms that search huge search spaces (environments) for potential solutions." (Belew, 1990, p44). Where the current work differs from Anderson (1990) is that it emphasizes the role of learning in the adaptive process, so as to address the development of memory through childhood and in the adult.

The objective of this thesis is to produce a preliminary account of memory phenomena in which the performance of the memory system is a consequence of the interaction of a learning mechanism and the environment in which the system is immersed. To provide an account that takes into consideration the environment, the mechanism and the interaction between them, the following three questions must be answered:

1. Can the environment relevant to particular phenomenon be adequately characterised?

2. Can a model which integrates learning into memory be constructed?

3. When the learning mechanism is considered in the context of the environmental statistics does it produce the behaviour that is found in the experimental setting?

Each of these questions is considered in turn.

To statistically characterise the environment it is necessary to record the items to which the memory system is exposed. As Anderson (1990) notes, such a characterisation is, in practice, difficult to achieve. The approach adopted has been to consider only written text (allowing the analysis of electronic text repositories), and to segment it at word boundaries. While written text is only a small fraction of the stimuli with which the memory system is presented, written words are often the materials used in memory experiments. If an adaptive mechanism is in operation one would expect that performance on written materials would be most closely linked to the written environment.

To address the second question one must propose and simulate a model of memory which learns. In doing so there are a number of issues which must be considered. What aspects of memory should the model learn? In chapter three I suggest that the three main candidates are the representations on which the model operates, the decision criteria and the control paradigms. In addition the model must retain the important characteristics of current memory models, such as the low degrees of interference in recognition and the ability to generalise as the vocabulary increases. To answer the second question in the affirmative the model should address all of these issues simultaneously.

It is possible to construct a learning mechanism that, while able to adapt to many environmental contingencies, is insensitive to the statistics that are relevant to a particular phenomenon. Further, there may be components of the memory system that are not adaptive to any great degree, and in these cases an accurate model would be insensitive to certain statistics. One might also produce a model that was sensitive to the correct statistics, but that produced behaviour that differed from that observed in the laboratory. To verify that the performance of an interactive model matches the behavioural data, a training set that embodies the environmental statistics must be constructed and the learning architecture must be applied to this training set.

In this section I have endeavoured to establish the motivation for the thesis so as to provide a context for the reader. However, when scientists approach their work they bring with them a broader context - the framework which embodies their set of professional prejudices (Hull, 1988). In order to place the work to be presented in

perspective it is useful to establish the framework from which it is being presented. In particular, it is useful to examine what type of constraints are to be considered and with what gravity. It is to this task which I now turn.

Constraints in the Cognitive Paradigm

Many constraints have been brought to bear in the research of cognitive functioning. To the extent that cognitive psychology and artificial intelligence attempt to address the same questions, the major distinction is their relationship to *behavioural data*. While cognitive psychologists are prepared to limit their domain of concern in order to remain faithful (preferably in a quantitative fashion) to the data, practitioners of artificial intelligence have concentrated on the principles underlying a variety of the functions of cognition. Collins and Smith (1988) claim that cognitive science lies at the interface of artificial intelligence and cognitive psychology. The aim is to uncover principles of cognition while retaining some connection with the behavioural data. The Hebbian Recurrent Network (HRN) introduced in chapter three attempts to fulfill both these criteria. It illuminates principles such as the learning of representation, decision criteria and control paradigms, but, is also expected to capture the behavioural data in a qualitative fashion. It is not intended as a detailed psychological model of memory phenomena.

The danger one faces as one moves away from behavioural data is that one's models become irrefutable. In place of the empirical constraints, however, cognitive science adds a number of *computational constraints*[2]. A cognitive model should be able to be implemented as an algorithm and the time and memory requirements should be within the bounds of reason, generally polynomial rather than exponential in the size of the input. Hence, the HRN architecture is an existence proof that a model of memory which learns can be implemented in a computationally feasible way.

In summary, the major constraints which have played a role in the work to follow are behavioural, computational and environmental.

[2]Marr (1982) defines a computational level which is concerned primarily with *what* computation a device performs and *why* it performs this computation. The computational constraints to which I refer, are that a model is implementable as an algorithm and that the time, memory and number of processors required by the algorithm are at worst polynomial. These constraints operate at the algorithmic/representational level of Marr's hierarchy.

Chapter 1

Introduction: Learning in Human Memory

1.1 The Status of Learning in Models of Human Memory

In the last two decades the idea that the human cognitive system is a serial digital computer has dominated cognitive research (Bates & Elman, 1992). This, *information processing (IP)*, approach has replaced the notion that the cognitive architecture is a consequence of the *interaction* between a genetically constrained learning system and the environment to which it is exposed[1]. Instead, the emphasis has been on the data structures and access processes which underpin cognition. In such a system the environment may "trigger" the onset of *maturation* events but does not play a causal role in determining the nature of the cognitive architecture.

A number of factors contributed to the demise of learning theory (see Bates & Elman, 1992, for a review). Among these was the undermining of some of the tenets of the leading proponent of an interactivist approach, Jean Piaget (Piaget, 1970a, 1970b, 1971; Piaget & Inhelder, 1973). In particular, it became clear that there were many more innate sensorimotor "schemes" than Piaget had originally thought which lent credence to the idea that most of the cognitive architecture was in place at birth (Bates & Elman, 1992). Also, the stages through which he proposed that

[1] c.f. Piaget's genetic epistemology.

development proceeded were found to be incorrect in detail (Bates & Elman, 1992). Further, processes such as accommodation and assimilation, which Piaget invoked to explain how change occurred in the cognitive system, were not as well defined as processes such as searching or sorting which were available in the information processing paradigm (Bates & Elman, 1992). In tandem with the development of artificial intelligence, learning came to mean the building of more complex structures from the piecewise composition of primitive representations. Within such a framework, however, there is no way to acquire the primitive representations (see Fodor, 1980, and appendix A) which lead to the conclusion that much of the content of the cognitive system must be innate, as is suggested in the following quote by Charniak and McDermott (1985):

> "One idea that has fascinated the Western mind is that there is a general purpose learning mechanism that accounts for almost all of the state of an adult human being. According to this idea, people are born knowing very little, and absorb almost everything by way of this general learner.
> ...
> We believe this idea is dead, killed off by research in AI (and linguistics, and other branches of "cognitive science"). What this research has revealed is that for an organism to learn anything, it must already know a lot. Learning begins with organised knowledge, which grows and becomes better organised. Without strong clues to what is to be learned, nothing will get learned."

(Charniak & McDermott, 1985, p 609)

The rejection of learning as an important aspect of the cognitive system is taken to its logical extreme by Piatelli-Palmarini (1989) who suggests that the term learning should be expunged from scientific discussion completely:

> "... I, for one, see no advantage in the preservation of the term learning. We agree with those who maintain that we would gain in clarity if the scientific use of the term were simply discontinued."

(Piatelli-Palmarini, 1989, p 2)

In the late eighties and early nineties, however, the role of learning in the cognitive system is being re-examined. Connectionism has answered many of the challenges that lead to the original demise of learning theory. For instance, connectionist models provide a well defined notion of learning[2] which is as readily implementable on current computers as the processes of searching or sorting. Further, the representations in the hidden layers of multi-layer perceptrons are not formed by the piecewise composition of primitive representational atoms, but emerge from a non-representational substrate in response to the environment. Hence, connectionist models can circumvent the objections to strong learning raised by Fodor (1980, see appendix A).

To be sure, there are still significant issues which connectionism has to face. Fodor and Pylyshyn (1988) have argued that connectionist systems do not display the same systematic use of representation observed in humans. There have been numerous attempts to address the objection and the discussion has yet to be concluded (Smolensky, 1988; Pollack, 1988; Brousse & Smolensky, 1989; Chalmers, 1990; Hadley, 1993). What has become clear is that connectionist models are capable of learning to process complex representations (Pollack, 1988; Chalmers, 1990) and to generalise this processing to massive numbers of novel examples (Brousse & Smolensky, 1989; Phillips & Wiles, 1993), two important properties of a theory of cognition.

As in other areas of cognitive science, the area of memory has seen a progression of models in which learning has played a very limited role. Despite substantial evidence that there is a great deal of change which takes place in subjects' ability to perform memory tasks both during development and as the adult gains experience in a certain memory paradigm, most current models deal primarily with the storage and retrieval operations and do not address how these processes may have been formed.

However, learning has not always been a secondary concern. During the late 1960s and early 1970s there was a move away from theories of verbal learning which had been inspired by the animal learning literature and which were based on multi-trial stimulus/response paradigms (i.e. *learning-to-learn* paradigms). They were

[2]The analysis of learning as descent on an energy gradient (e.g., Hopfield, 1982, 1984) has aided the mathematical formalisation of the process.

replaced by theories which were based on the information processing approach and which placed greater emphasis on single trial retention of familiar items. The shift is chronicled in the volumes of the "Annual Review of Psychology" between 1965 and 1975. Blough and Millward (1965) entitled their review "Learning: Operant Conditioning and Verbal Learning", clearly demonstrating the ties to the animal learning literature. Keppel (1968), although less explicit with respect to the links to the animal literature, was primarily concerned with verbal learning issues such as response learning, associative learning and transfer. He also heralded the change that was taking place, including a section on memory that covered topics such as storage, interference and decay.

Tulving and Madigan (1970), in their critique of the state of the art, noted that the community had divided into the "students of verbal learning" and the "students of memory". Verbal learning was about the "acquisition and retention of verbal responses to stimuli" (Tulving & Madigan, 1970, p 439) while memory was about the "storage and retention of item and order information" (p 440, Tulving & Madigan, 1970). In 1973, Greeno and Bjork (1973) completely omitted reference to verbal learning and were explicitly influenced by work in artificial intelligence. They entitled their review "Mathematical Learning Theory and the New Mental Forestry" to highlight the importance of the tree data structure which had been borrowed from computer science. Perhaps the most telling indication that the "students of memory" were successful came in 1975 when Postman (1975) made no reference to the verbal learning literature, despite the fact that he had been one of the major contributors to the area only five years earlier.

Estes (1991), in his review of human memory, argues that learning is a major aspect of the memory system and foreshadows that connectionism will renew the search for models in which the processes of memory are formed:

> "The special relevance of connectionist models to my long term research interests and to this essay is that they have offered the promise of rescuing learning theory from its near eclipse during several decades in which cognitive research has been dominated by concern with problems of information-processing, representation in memory, and cognitive operations. Whether we look at research and theory on short-term memory,

> semantic memory, propositional networks, or even the earlier distributed memory models, we see almost no efforts to interpret in any detail the processes whereby information comes to be stored in memory. ... Thus in a connectionist approach, learning is not viewed as a subsidiary problem to be left for the consideration after the big problems are taken care of, but, rather, is a major aspect of the cognitive system from the start. For an investigator who grew up scientifically in the golden age of learning theory, this renewed emphasis on learning is a welcome development."

(Estes, 1991, p 15)

The challenge which Estes (1991) poses is to incorporate learning into models of human memory and in this thesis I take up this challenge by investigating models of recognition memory which learn. I start by defining learning and reviewing current models of recognition memory. This is followed by a section discussing why learning is an important aspect of the memory system. Of the reasons presented in this section it is the relationship between the environment and the mechanism which will be the focus of the thesis. Hence the next section expands upon this relationship. The chapter concludes with an outline of the thesis.

1.2 Learning Defined

The idea of learning has been contentious for several decades. At least part of this debate has revolved around the definition of the term and hence in this section I develop a working definition.

Perhaps the broadest definition, and the one that is in colloquial use, links learning to a change in behaviour. A subject is considered to have learned if their subsequent behaviour is altered due to exposure to an environmental stimulus[3]. Inherent in this definition is the idea that learning is the consequence of the interaction of a learning system and the environment.

Experience in *learning-to-learn* paradigms suggests a division of environmentally evoked behavioural change into storage and learning components. Exposure to a

[3]Further, there seems to be a sense that this change should be derived internally. For instance, having a leg amputated certainly involves changes in behaviour due to environmental conditions but would not be considered learning since the alterations were derived externally.

single trial leads to the storage of information which is then available for subsequent retrieval while exposure to multiple trials leads to the gradual improvement in performance on a task. According to Postman, "learning to learn has been taken to denote the acquisition of higher-order habits and skills which produce more or less permanent changes in the subject's mode of attack on the task" (p 242, Postman, 1969). This definition reflects my focus on the acquisition of the skills of memory, that is, on the learning to learn component.

In order to implement this notion of learning it is necessary to identify what is changing as "skills" and "habits" are acquired. The most likely candidates are the control processes (Atkinson & Shiffrin, 1968) including encoding procedures, rehearsal operations and search strategies. It is these processes which show a marked developmental course (Hasher & Zacks, 1979) and the most variation, both between individuals and between cultures (Kail, 1984).

However, the description of a process also requires determining the representations on which it operates. In drawing a distinction between learning and storage, it also necessary to distinguish the process of forming representations from that of constructing the specific traces that will be entered into memory[4]. The nature of these representations has been a source of much debate in the literature (Shiffrin & Murnane, 1991). A learning system in which representations emerge to fulfill the requirements of the task may provide a resolution to the argument.

In the memory literature, the decision criteria used in recall and recognition, which would normally be considered part of the control processes, are often considered as separate components. However, these might also be learned and it is useful to consider representations, decision criteria and control processes separately. Hence, I propose the following definition:

> **Learning:** A change in behaviour due to the formation of representations, decision criteria and control processes as a consequence of the interaction of a mechanism and the environment.

The definition is used in the next section to review the learning characteristics of current mathematical models of human memory.

[4]Thanks to an anonymous reviewer for emphasizing this point.

1.3 Review of Current Models of Recognition Memory

In this section, a representative set of recent recognition memory models will be reviewed with respect to their learning capabilities. In a typical recognition experiment, subjects are presented with a list of items or pairs of items (usually words) which they are required to study. The study list is followed by a test list (often after a retention interval) which again contains items or pairs of items some of which did occur in the study list and some which did not. Subjects are asked to respond either "old" indicating that the test item or pair was in the list, or "new" indicating that it was not. While there are many variations to this study/test paradigm, the models which are reviewed all aim to address the task which has just been described.

The global matching models (Humphreys, Pike, Bain, & Tehan, 1989) including TODAM (Murdock, 1982), SAM (Gillund & Shiffrin, 1984), Minerva II (Hintzman, 1984), and the matrix model (Pike, 1984; Humphreys et al., 1989) have dominated the literature in the last decade and hence will be considered first. In addition, the review will include the backpropagation based model studied by Ratcliff (1990) as it displays some learning characteristics and will form part of the model which will be presented in chapter three. In keeping with the definition of learning presented above, the following sections summarise each model in terms of their representations, decision criteria and control processes.

1.3.1 Theory of Distributed Associative Memory

The Theory of Distributed Associative Memory (TODAM, Murdock, 1982) exemplifies many ideas which have influenced work in the area of human memory during the last decade. The major objective of TODAM was to demonstrate how both item and associative information could be stored in, and retrieved from, a distributed memory system[5].

[5]The Composite Holographic Associative Recall Model (CHARM, Eich, 1982), although addressing different issues, is very similar to TODAM in mechanism, and hence has not been reviewed separately.

Representations: Items are represented by random vectors. The components of these vectors are chosen from a normal distribution ($\mu = 0$ and $\sigma^2 = P/N$ where P is the power of the vector and N is its length). Associations are constructed by convolving the items. The memory is a single vector to which both item and association vectors are added.

Decision Criteria: The decision criterion for recognition is based on signal detection theory (Swets & Green, 1961). A familiarity value is calculated by taking the dot product of the memory vector with the cue vector, adding a time varying noise component and comparing the result against two criteria. If the familiarity is above the higher criterion a "yes" response is given. If the familiarity is below the lower criterion a "no" response is given. If the familiarity falls between the criteria then the system waits until the noise takes the familiarity value beyond one of the criteria.

Control Processes: The control process can be divided into the study and test phases. During study, items and/or pairs are presented to the memory system. Each item is added directly to the memory vector and for each pair of items the convolved association vector is added. In the test phase either items or pairs of items are presented. The decision criterion is applied and the system waits for a response.

In TODAM there is no sense in which the representations, decision criteria or control processes are a consequence of the environment. They are all predefined. The only component which does change with the environmental input is the memory vector.

1.3.2 Search of Associative Memory

The Search of Associative Memory (SAM, Gillund & Shiffrin, 1984) model of recognition was introduced to describe the relationship between recall and recognition. In particular, Gillund and Shiffrin (1984) recognised that subjects were able to make very fast negative judgements in recognition. If these judgements were the result of an unsuccessful search process as had been suggested by (Raaijmakers & Shiffrin, 1981), then latencies for negative decisions should have been significantly longer

than for positive items. Hence, they added a global recognition procedure to the earlier SAM model of recall to account for the fast negative latencies as well as a number of other results from recall and recognition paradigms.

Representations: Gillund and Shiffrin (1984) assume that the representations in long term store are "closely interconnected, relatively unitised, permanent sets of features called *images*." (p 7 Gillund & Shiffrin, 1984). These images contain elements which identify the context in which they occur, item information including its name, and interitem information (Gillund & Shiffrin, 1984). The internal structure of the image representations is not specified, but for the purposes of the simulations each image corresponds to a unique element of a vector. The associations between cues and images learned during the study phase are represented by the connection *strengths* in a matrix called the *retrieval structure*. These include the links from other item cues to the images as well as those from the context cue.

Decision Criteria: As in TODAM, a global familiarity value is calculated. The strength of the context to image association is multiplied by the item cue to image strength for each of the images in memory. These values are then added to produce the familiarity value which is compared against a criterion. If it is above criterion a "yes" response is generated. If it is below criterion a "no" response is generated. Uncertainty is introduced by adding a noise component to each of the strengths.

Control Processes: After the retrieval structure has been filled as a consequence of exposure to the study list[6], the cues available to the system at test are assembled in short-term memory, and the familiarity value is calculated and compared against the decision threshold (see figure 5, Gillund and Shiffrin (1984) for a flowchart of the test process).

As with TODAM, the only components of SAM which are altered by the environment are the associative (or connection) strengths within the retrieval structure.

[6]The steps involved in filling the retrieval structure are not outlined in any detail by Gillund and Shiffrin (1984), however it is assumed that the strengths in the retrieval structure are related to the length of time that the items are in the short term memory during study.

1.3.3 The Matrix Model

The matrix model (Pike, 1984) was introduced to address similar issues as TODAM, but used matrix operations rather than convolution/correlation to simplify the mathematics and decrease the amount of noise. Later it was extended from rank two to rank three tensors to allow the storage of three way information between the items of a pair and the context (Humphreys et al., 1989). Such information is necessary to account for cued recall paradigms such as the crossed associates (Spence & Lair, 1965; McCullers, 1965) and the ABABr paradigm (Jung, 1962; Barnes & Underwood, 1959) in which two way associations are not sufficient (see chapter two for an explanation of these paradigms).

Representation: As in TODAM, items are represented by random vectors in the matrix model. Each of the components is Gaussian with mean $1/\sqrt{n}$, and variance σ^2 where n is the number of components of the vector. Associations are constructed by taking the outer product of the items to be associated. In single item recognition these are the item and the context. These associations are then added to create the matrix memory of the study list.

Decision Processes: Again signal detection theory is applied. The familiarity measure is calculated by multiplying the matrix memory by the context to get a vector which is the addition of the list items. In a similar fashion to TODAM, the dot product of the cue item and the result vector is then determined and compared against a criterion.

Control Processes: During the study phase, the outer product of the study items and the current context is formed and added to memory. At test, for the simple recognition task described above, the memory is multiplied by the context cue and the decision criterion is applied. The matrix model has been applied to many other retrieval tasks requiring a variety of control processes (see figures 1 and 2 in Humphreys et al. (1989) for an outline).

While the matrix model (Pike, 1984; Humphreys et al., 1989) has none of the learning characteristics mentioned above the rank two version can be thought of as a Hebbian network and can be easily integrated into an error correcting connectionist

network (Wiles, Humphreys, Bain, & Dennis, 1990). Hence, it serves as the storage mechanism of the Hebbian Recurrent Network (HRN) which is presented in chapter three.

1.3.4 Minerva II

When subjects are exposed to a single list in which an item is repeated, they are able to judge the serial position of both occurrences of the item (Hintzman & Block, 1971). This finding lead Hintzman (1984) to propose the Minerva II model of memory in which each presentation of an item was entered into memory as a separate trace and, hence, could be extracted individually. He argues that models which sum trace strengths are not able to distinguish between the individual traces and, hence, cannot judge the position of repeated traces.

Representation: As with TODAM and the matrix model, items are represented in Minerva II by vectors. Each vector element is termed a "feature loading" and will be zero if the feature is irrelevant or unknown, one if it is present and negative one if it is not. Associative information is stored by concatenating the vectors for different items. Rather than creating a composite memory, the traces in Minerva II remain separate and information is integrated at retrieval time.

Decision Criteria: The first step in calculating the familiarity value of a probe item is to calculate its similarity to each of the traces in memory. The dot product of the probe and the trace is divided by the number of components for which either the probe or the target is non zero. Hence, the similarity is one if the probe and trace are identical. Each of these similarity values is raised to the third power to give an activation value for the trace. The activation values are then added across the traces to produce the familiarity value. It is this familiarity value which is compared against the criterion.

Control Processes: As each item is presented in the study phase, it is probabilistically encoded and becomes a new entry to the set of memory traces. When the probe is presented, the decision criterion is applied directly to the trace set.

In addition to memory paradigms such as paired associate recall, frequency

judgement and recognition (Hintzman, 1984, 1988), Minerva II has been used to model tasks in concept learning (Hintzman, 1986). Unlike the theorists examined thus far, Hintzman makes his nativist stance explicit. In his view, concepts are represented by the activation of a large set of primitive atoms which are not acquired through experience[7].

> "Every experience is assumed to be represented internally as an active configuration of primitive properties... The number of primitive properties is assumed to be large and to range from simple emotional tones and modality-specific sensory features (e.g. basic colours and odours) through properties accessible by more than one modality (e.g. intermittency, spatial location), to primitive abstract relations (e.g. before, same as, greater than, has as parts). These properties are not acquired by experience, although the ability to label them may be.
>
> In assuming a large set of primitive properties of which some are abstract, the present theory takes a rationalist, as opposed to empiricist, view."

(Hintzman, 1986, p 412)

Hintzman's definition of concept learning is restricted to the assembling of the primitive features (some of which may encode the name of the item) into item vectors and the abstraction of schema from these vectors. In contrast, learning as defined in this thesis assumes that representations emerge from a largely contentfree substrate (see appendix A). Hence, learning would include the ability to form representations for primitive abstract relations and cross-modal properties. In this sense, Minerva II does not satisfy the criterion of learning representation. Nor does Minerva II acquire the decision criteria or control processes with which it operates.

1.3.5 The Multilayer Backpropagation Network

Multi-layer connectionist models of recognition memory were evaluated by Ratcliff (1990). While error correcting connectionist models had been applied successfully

[7] Fodor (1980, 1981) argued that such a position is necessary to account for the concept learning literature.

in a number of domains within cognitive science there had been no systematic application to memory phenomena. Furthermore, in Ratcliff's opinion, there was an incomplete understanding of the limitations of such models and his objective was to gain a greater appreciation of these constraints by using a three layer backpropagation network to model basic interference data in a quantitative fashion.

Representation: In the network described by Ratcliff (1990), items were represented twice. Firstly, the network was presented with vectors consisting of zeros and ones. In one simulation these were chosen so that the items were orthogonal (using local codes) and in another they were chosen randomly with equal probability of a zero or one. The network was required to reproduce its input and hence the vectors used as targets were identical to the input vectors. In addition, items were represented again on the hidden units. This representation is formed by the learning mechanism as a result of exposure to the environment and, hence, does satisfy the criterion of learning of representation specified above.

Decision Criteria: In the initial architecture explored by Ratcliff (1990), a dot product of the reconstructed output and the input was the familiarity measure. Just as in TODAM, SAM, Minerva II and the matrix model, this familiarity measure was compared against a criterion to determine the output. Such a decision criterion is not learned. However, in a further section on alternative architectures, he discusses the affect of adding a "Yes node" (as an output), which is expected to be near one if the item has been studied, and near zero if not. For such an alteration to function, however, the network needs to be trained on both positive and negative examples. Ratcliff (1990) argued that during the study phase of typical recognition experiments, there is little evidence that people are training on negative examples. When he simulated the network, without the inclusion of negative examples, performance was very poor and hence he concluded that backpropagation networks are not well suited to modelling recognition performance. However, such an architecture is interesting, in that, the new/old decision is made by the network as a consequence of training. The difficulty with the training only occurs if one attempts to use the network as a storage mechanism (see chapter three).

Control Processes: As in the previous models the control of the backpropagation model is chosen a priori. During the study phase, items are presented individually and trained to criterion. During test, each item is presented to the network, the output calculated and the decision mechanism applied. Again there is no sense in which the control aspects of the model could be said to have been learned.

As a model of human recognition performance the multilayer perceptron is deficient in a number of ways (Ratcliff, 1990). It suffers from massive retroactive interference and does not show a monotonic improvement in distinguishing new from old items as a function of training. While there have been attempts to address the difficulties (see chapter three for a review of these attempts), the problem has its roots in the underdeveloped role of learning (which backpropagation is good at) as opposed to memorisation (which backpropagation is bad at) in current models of memory.

In all of the models presented, learning is seen as a subproblem to be dealt with once the "big" problems are explained (Estes, 1991). Not only does this perspective underestimate the importance of learning in explaining a range of memory phenomena, but it fails to identify the constraints that learning can place on memory schemes. Some of these are outlined in the next section.

1.4 Why learning is important in memory

There are many reasons why learning is important to a theory of memory. The adult memory system is the product of a developmental process. The inclusion of a learning mechanism into models of human memory introduces the possibility of addressing how the representations, decision criteria and control process which are assumed in current mathematical models might have developed (see section 1.4.1). Indeed, the development of memory skills is not limited to childhood. There is an extensive empirical literature on learning-to-learn effects (see Postman (1969) for a review) which can also start to be addressed with the addition of a learning mechanism (see section 1.4.2). Furthermore, memory is a component of the cognitive system. The philosophical debate on learning is as relevant in memory as it is to other cognitive domains such as language (see section 1.4.3) and the introduction of learning provides a cross paradigm generality not available to traditional informa-

tion processing approaches (see section 1.4.4). Finally, as outlined in the preface, the inclusion of a learning mechanism encourages a more dynamic role for the environment (see section 1.4.5). In the following subsections I will expand on these points ending with the role of the environment which will be the focus of the next section.

1.4.1 Learning and Development

Connectionist models have been successful in explaining several aspects of the development of human cognition (see Plunkett & Sinha, 1992; Bates & Elman, 1992; Shultz, 1991, for reviews). By virtue of their learning mechanisms, connectionist models are equipped to model the change which takes place as the child grows in their experience of the world (Bates & Elman, 1992). The developmental literature demonstrates that memory skills such as imaging, elaboration, naming and rehearsal are learned (Hasher & Zacks, 1979; Kail, 1984), but connectionist models have yet to be applied in this domain.

Any such application will have to take into account domain specific structural dispositions (Karmiloff-Smith, 1992a, 1992b, 1992c). Learning to retain information to the extent observed in memory tasks is computationally intensive and perhaps untenable. Hence, in this thesis a memory specific predisposition to store information is added to a model in which the encoding and retrieval processes are implemented in an optimising connectionist fashion.

1.4.2 Learning-to-Learn

Change in the memory system is not restricted to childhood. Under a broad range of conditions adult subjects improve their performance on memory tasks as they gain experience. During the sixties and early seventies there was a great deal of work on learning-to-learn beginning with studies by Duncan (1960), who promoted learning-to-learn as a subject of study in human verbal learning[8]. Postman, Keppel and their colleagues produced a series of studies which investigated empirically the conditions that influence learning-to-learn. These studies varied the similar-

[8]Learning-to-learn was already a well established phenomenon in the animal literature and had been noted as an ancillary finding in some human studies. See also McGeoch and Irion (1952).

ity relations between successive tasks (Postman & Schwartz, 1964; Postman, 1964; Keppel & Postman, 1966; Postman, 1968a), the paradigms such as paired associate learning, serial recall and free recall (Postman & Stark, 1967; Postman, 1968b; Postman et al., 1970; Martin, Simon, & Ditrichs, 1966) and explored other constructs such as warm-up (Schwenn & Postman, 1967), associative practice (Schwenn & Postman, 1967), response integration (Postman, Keppel, & Zacks, 1968) and the interaction with proactive interference (McVaugh & Postman, 1969). In addition, there have been studies outlining the affects of stimulus qualities such as vividness (Tulving, McNulty, & Ozier, 1965) and meaningfulness (Underwood & Schulz, 1960) on learning-to-learn curves. Postman (1969) and Keppel (1968) provide summaries of much of this work.

The consequences to performance of previous experience (i.e. *transfer*) in learning-to-learn phenomena can be divided into two sources (Postman, 1969). The first is termed *specific transfer* and refers to effects that depend on the similarity relations of successive tasks. For instance, if subjects are exposed to a list of pairs of items and then a second list in which the second item of each pair has been replaced by a synonym (the ABAB' paradigm), they will perform more poorly on the second list than if the two lists were unrelated. As they are exposed to additional examples, however, this effect reverses and they become better on the synonyms list than the control list. Similar specific transfer is seen in other paradigms such as ABAC[9] and ABABr[10].

The second source of facilitation observed in list learning experiments is called *nonspecific transfer* and refers to the higher order transfer of skills that is independent of the similarity relations between successive tasks. For example, Postman et al. (1968) showed that subjects presented with number/trigram pairs showed a marked improvement in cued recall as a function of training. Underwood and Schulz (1960) divide nonspecific transfer into two components. The first is *response learning* which refers to the improvement due to the subject becoming more familiar with the items that are required as responses. For example, in the number/trigram experiment

[9] In the ABAC paradigm subjects are presented with a list of item pairs, followed by a second list in which the second items have been replaced by new items.

[10] In the ABABr paradigm the pairs from the first list are reorganised to form a second list. The same items occur in the second but with different partners.

some of the facilitation is taken to occur as the subjects became more proficient at responding with trigrams. The second form of facilitation is *associative learning*. To complete the cued recall task the subject must not only become familiar with the responses but must also be able to form associations to these responses. Postman (1969) suggests that this process is also learned.

Current mechanisms of memory do not address nonspecific transfer (see Murdock, 1987, for an exception). Both the response identification and association formation processes are considered fixed. While it is generally not difficult to imagine how the response identification may be learned (e.g. Chappell, 1993; Chappell & Humphreys, in press, 1993), the association formation processes are fundamental to each of the models and it is difficult to see how they could be extended to incorporate learning of the association mechanism without significantly altering the model.

1.4.3 Learning and Meaning

Memory is a part of cognition. The representations used in models of memory are often linguistic in origin and how they attain their meaning impacts upon the study of memory. In current models of memory a one to one correspondence between the objects in the world (usually words) and memorial representations is assumed. When relational information must be stored, such as in pair recognition, some form of functional composition operation (e.g. convolution, matrix multiplication) is often employed. However, Bickhard (1991) has argued that such a system is not feasible since it is not able in principle to account for the development of meaning. Appendix A provides a discussion of Bickhard's argument.

1.4.4 Learning and Cross Paradigm Generality

The ability to use the same architecture across a variety of domains (i.e. *cross paradigm generality*) is another advantage of introducing learning into models of memory. At the current time cognitive psychology has been divided into a number of subdisciplines each of which proceed with a great deal of autonomy. For instance, the study of memory phenomena is affected only mildly by the study of problem solving, categorisation or inferential reasoning. Similarly, processes such as percep-

tion and attention are investigated fairly independently of any constraints which empirical work in memory might raise. The autonomous nature of the various disciplines of cognition encourages a diversity of mechanisms designed specifically to solve the problems of individual areas. These area specific mechanisms will have to be integrated to form a cohesive story of cognition (Karmiloff-Smith, 1992c; Feldman, Lakoff, Stolcke, & Hollbach-Weber, 1990).

A learning mechanism provides a useful tool in the pursuit of such an integration. Consider for instance, a standard backpropagation network. The same network can at one time be performing categorisation, at another a memory task and in yet another case be involved in inferential reasoning. While there will be the need for domain specific constraint (Karmiloff-Smith, 1992c), error correcting connectionist networks provide mechanisms for bringing the components together (c.f. connectionist "glue" Waibel, 1989).

1.4.5 Learning and the Environment

The final reason for advocating the addition of learning into models of human memory is the expanded role which this allows for the environment. In a system that learns, the environment is not simply the source of inputs and the destination of outputs. It is responsible for the nature of the information processing mechanism. As the interaction of mechanism and environment will form the basis of the thesis, the role of the environment in models of human memory is developed in the next section.

1.5 The Role of the Environment in Models of Human Memory

The idea that the human cognitive system is adapted to the environment has a long history (Anderson, 1990; Brunswik, 1956; Campbell, 1974; Shepard, 1987). However, the role that the environment might play in models of human memory has, until recently, had little impact on the construction of mathematical models. The obstacle has been the difficulty in characterising the adaptive mechanisms which mediate between the environment and the memory system.

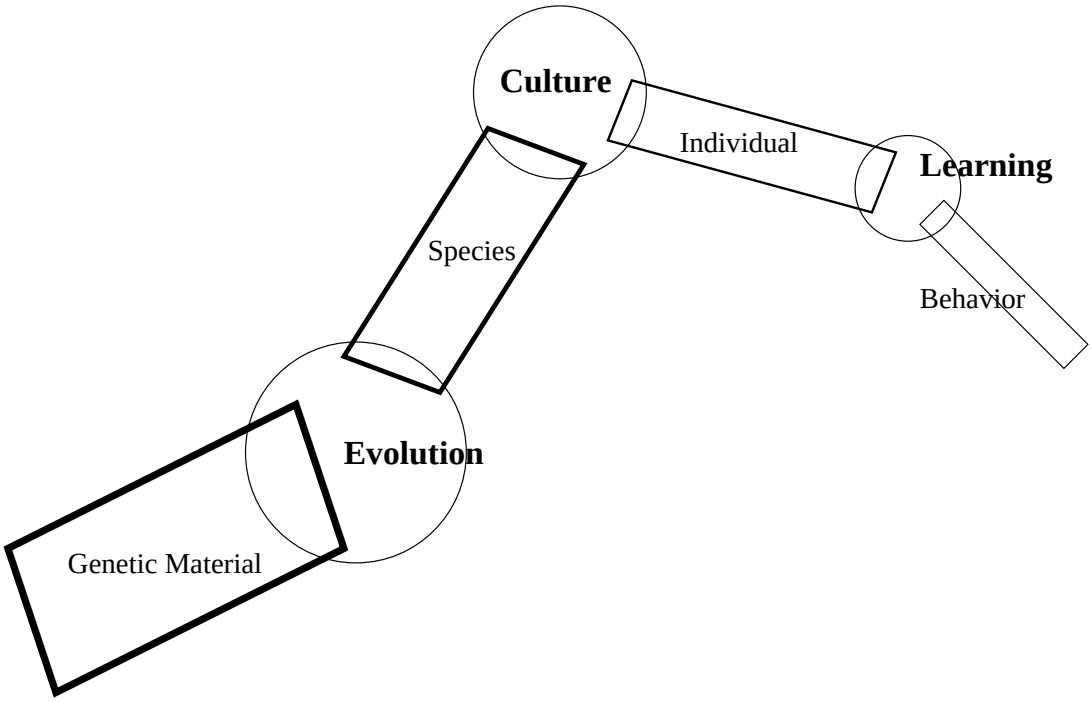

Figure 1.1: The robot arm framework for understanding the relationship between adaptive processes. The diagram illustrates the relationship between evolution, culture and learning. Evolution is the large pivot since it is the process which takes the longest period of time. Reconstructed from figure 19, Belew (1990).

The robot arm diagram (Belew, 1990, see figure 1.1) provides a framework for describing how the processes of evolution, culture and learning interact. They are seen as complementary mechanisms which serve to adapt behaviour to the environment. Evolution is the largest pivot to indicate that it has the most inertia and takes the longest to adapt. Culture is next and learning is the smallest indicating that it is the fastest of the processes.

Anderson's (1990), *rational analysis* is an explicit attempt to specify the contribution of evolution to the adaptive process and the next section outlines Anderson's approach. In this thesis, however, the focus will fall on the learning arm. Hence the following section outlines the *interactive* (or learning) approach and contrasts this with rational analysis.

1.5.1 Evolution and the Rational Analysis of Memory

Anderson and Milson's (1989)[11] rational theory is the first model of human memory to concentrate on the nature and affect of the environment. The basic premise is that the structure of the human memory system has evolved to meet the demands of the environment in which it is immersed and hence studying the structure of the environment should provide constraint on the types of mechanism which might be employed. In essence, it is an attempt to untangle the "why" of human memory rather than the "how", which is the dominant concern currently addressed in the literature, or the "what" which has recently been considered by Humphreys, Wiles, and Dennis (in press).

In applying the rational analysis framework to the area of human memory, Anderson and Milson (1989) start by framing the retrieval of a memory trace as the optimisation of an objective function.

Memories (A) will be retrieved until the following inequality holds:

$$p(A)G < C \tag{1.1}$$

where $p(A)$ is the probability that a memory will be required (*the need probability*), G is the gain in achieving the goal and C is the cost associated with retrieving each memory. G and C are the computational constraints. They are deliberately left undefined because Anderson and Milson (1989) are trying to determine what predictions they can make without having to commit to mechanistic assumptions. G is assumed to be related to the importance of the task. Hence, subjects will be inclined to try harder if a task has a greater importance. C, the cost of retrieving a memory, is presumed to be related to retrieval time but could also include factors such as metabolic expenditure. The variable which reflects the environment and, hence, the major variable to be estimated is the need probability of the trace, which is obtained by considering both the history of the memory trace and the available cues. These factors are then combined in a Bayesian fashion to produce the need probability. In the majority of the phenomena examined by Anderson and Milson (1989) the relationship of the need probability to latency and probability of recall is

[11] Anderson (1990) has applied the rational approach to a number of areas in cognition. See also Anderson (1991) for a review of the approach with peer commentary.

assumed to be monotonic. However, they also developed a more specific mechanism to allow quantitative predictions.

One of the major problems encountered by Anderson and Milson (1989) was the affect of subject strategy. While the strategy used by the subject has a marked affect on the memory performance, there is often little difference in the task that the system must perform or the optimal response to that task. Hence, they comment that subject strategies fall outside the scope of their evolutionary explanation[12]. To deal with this deficiency in the context of free recall, Anderson and Milson (1989) propose a hybrid mechanism, in which a rehearsal buffer is used in conjunction with the optimised memory. With these assumptions in place they then account for serial position data (Murdock, 1962) and the affects of the presentation time and list length (Roberts, 1972).

The relationship between environment and mechanism: What makes rational analysis different from traditional models of memory is that it employs an optimality principle (or teleological argument). A rational level of analysis which embodies the structure of the environment is assumed to overarch mechanistic models which have been optimised to this structure. Discussion of the environment/mechanism interface is restricted to the discovery of correspondences between components of the environment and components of the mechanism. For example, Anderson (1990) uses correspondence to describe how the ACT* model relates to the rational model:

> "... activation is the analogue of need probability ... Node strength ... is the analogue of the history factor in the rational theory. Link strength .. is the analog of the contextual factor in the rational analysis..."

(Anderson, 1990, p 83)

Figure 1.2 shows diagrammatically how components in the rational (or environmental) level *correspond* to components of the algorithmic (or mechanism) level.

However, Bickhard (1991) has argued that no adaptive mechanism, whether it be learning, culture or evolution could generate such a set of correspondences (see

[12] Yet, the environment is implicated in the development of strategies. For instance, people who do not under go formal schooling do not acquire the ability to rehearse (Wagner, 1978). What seems to be required are mechanisms which tune themselves to cultural or individual learning histories.

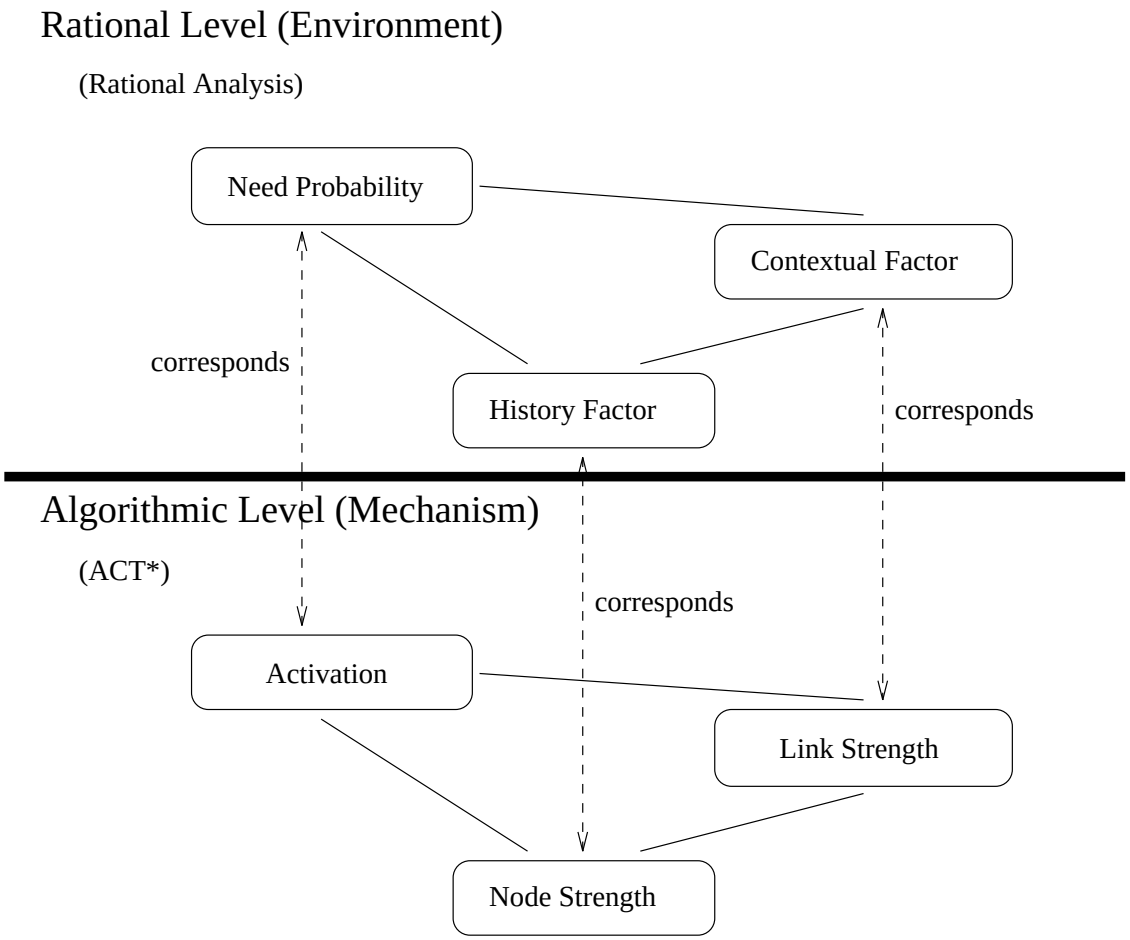

Figure 1.2: Correspondence relationships between rational analysis and ACT*. The diagram illustrates Anderson's (1990) proposal that the environment and the mechanism are best dealt with at different levels of theory. The task is to identify which components of the rational analysis of the environment correspond to which components of the model of the mechanism, in this case ACT*. The diagram illustrates the correspondences which Anderson finds.

appendix A) and that this undermines rational analysis (Campbell & Bickhard, 1991).

Alternatively representations could differentiate environments rather than correspond to them (i.e. *interactivism*, Bickhard, 1991). In such a system representations emerge in order to meet the requirements of the environment. They are not constructed from primitive elements which correspond to objects in the world. This is the approach adopted in this thesis and the next section explores the affect this has on the relationship between the environment and the mechanism.

1.5.2 Learning and the interactive approach

The memory literature covers an enormous range of phenomena (Humphreys et al., 1989; Tulving, 1983). Furthermore, these phenomena display a complex structure. Modellers of human memory have been faced with a dilemma. Either they try to build a model which covers a significant portion of the field relying on fairly imprecise verbal principles[13], or they build very specific mathematical models which require many quite arbitrary assumptions and apply only to a limited subset of the experimental literature [14].

It is my contention, however, that a significant portion of the complexity of human memory is a consequence of the interaction with a complex environment rather than the result of an inherently complex mechanism. Assuming a relatively simple learning mechanism allows the structure of the environment to be captured without incorporating it into the mechanism. In contrast to the correspondence relationship advocated in rational analysis, a learning model based on emergent representation allows the environment and the mechanism to assume complementary aspects of an explanation. Since the environment is observable, environmental assumptions tend to be more amenable to empirical verification (Anderson & Schooler, 1991; Dennis et al., 1992). If it is the case that experimental results are to some degree determined by the structure of the environment, a model which learns should provide a parsimonious way of accounting for the experimental literature.

The aim of this thesis (outlined in the next section) is to produce a demonstration of a memory model in which the performance is a consequence of the interaction of a learning mechanism and the environment.

1.6 Outline of the Thesis

The thesis is outlined in figure 1.3. As discussed in the preface it endeavours to answer three questions:

1. Can the environment relevant to a particular phenomenon be adequately characterised?

[13] See Hintzman (1991) for a criticism of this approach.
[14] See Watkins (1990) for a criticism of this approach.

2. Can a model which integrates learning into memory be constructed?

3. When the learning mechanism is considered in the context of the environmental statistics, does it produce the behaviour that is found in the experimental setting?

Chapter two addresses the first question by investigating empirically the environment of memory relevant to the word frequency and the null list strength effects in recognition. In addition, some of the difficulties encountered when trying to provide an environmental analysis of cued recall paradigms (such as the crossed associates paradigm and the ABABr paradigm) are discussed.

Chapter three presents a learning mechanism (c.f. question two) called the Hebbian Recurrent Network (HRN) which incorporates the matrix model of Pike (1984) into the Simple Recurrent Network (SRN) of Elman (1989, 1990), and outlines a series of simulations which establish its memory and learning characteristics.

Chapter four completes the interactive picture by applying the HRN in light of the environmental analysis of chapter two to model the word frequency effect in recognition (c.f. question three). Chapter five provides a second example of the interactive approach by presenting an account of the null list strength effect in recognition.

The conclusions and further work chapter summarises the contribution, evaluates the approach and outlines some possible extensions of the work.

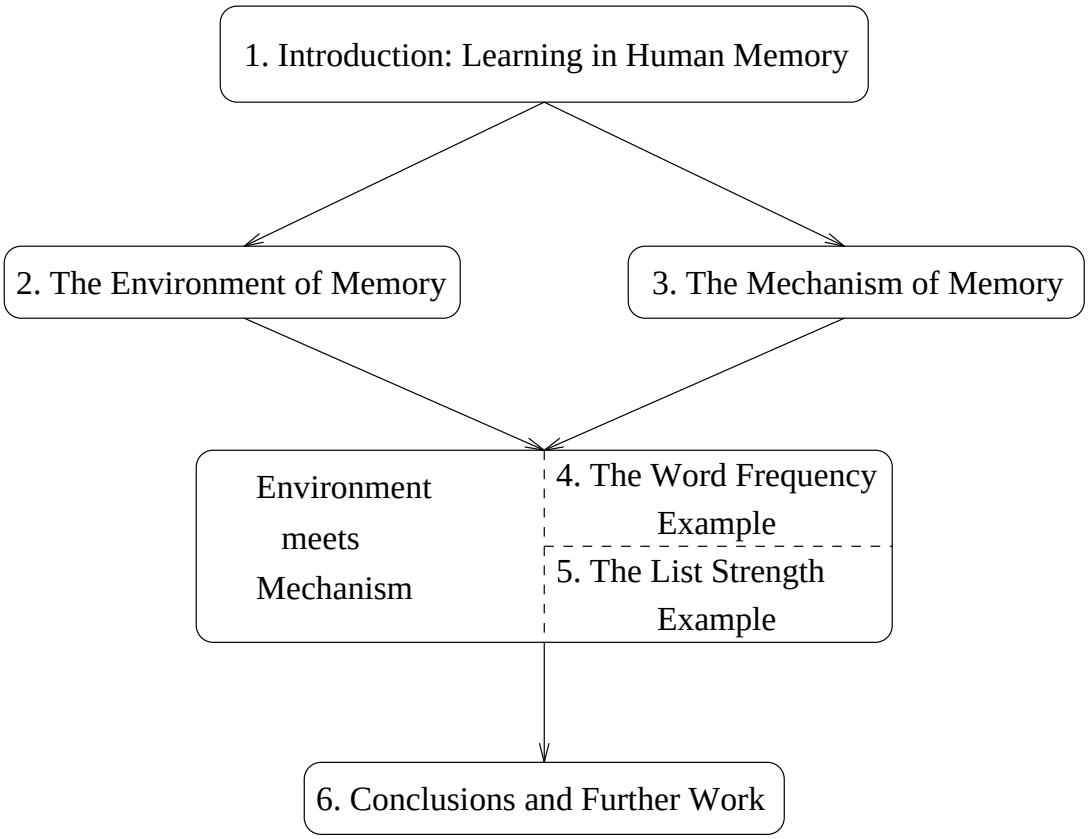

Figure 1.3: Map of the thesis.

Chapter 2

The Environment of Memory

In this chapter, the hypothesis that the memory that is observed in the laboratory is a consequence of both an adaptive mechanism and the environment to which that mechanism has been exposed, is explored. The primary methodological implication of such a hypothesis is that memory cannot be understood without an empirical appreciation of the environment (Anderson, 1990). Anderson and Schooler (1991) provide the first such formal analysis of the environment of memory. The aim of this chapter is to provide additional analyses for two phenomena that provide strong constraints on models of memory, namely, the affect of word frequency on recognition (Gorman, 1961; Schulman & Lovelace, 1970) and the null list strength effect in recognition (Shiffrin, Ratcliff, & Clark, 1990; Ratcliff, Clark, & Shiffrin, 1990). The word frequency effect is one of the clearest examples in established research of the environment impacting upon memory performance and, hence, is a natural place to begin the analysis. The null list strength effect, which has received a significant amount of attention in recent years (Murnane & Shiffrin, 1991a; Ratcliff et al., 1990; Yonelinas, Hockley, & Murdock, 1992; Shiffrin et al., 1990; Murdock & Kahana, 1992; Chappell & Humphreys, in press, 1993; Chappell, 1993; Heathcote, 1993), is not so obviously contingent on the environment and, therefore, provides a test of the generality of environmental analysis.

When undertaking a study of the environment there are at least two possible approaches. The first is to build a model of the environment and then demonstrate its validity. Essentially, this is the approach adopted by Anderson (1990). Such a methodology has lead to criticism from commentators such as Gigerenzer (1991)

who argue that Anderson "seems to have started with Bayes' theorem as a model of rationality and to have assumed that the structural assumptions underlying Bayes' theorem specify the structure of the environment as well." (p 496, Gigerenzer, 1991). However, the environment can be studied independently of any Bayesian assumptions. For instance, experimental phenomena could be considered on a case by case basis by identifying the everyday tasks in which they are involved and collecting the appropriate statistics. As indicated in the previous paragraph, this is the approach that I will adopt.

In the sections that follow, the chosen phenomena are explored in two parts. The first part reviews the experimental data and current accounts and the next determines the environmental statistics relevant to the phenomena. These sections are followed by a brief discussion of the issues involved in extending the analysis to cued recall paradigms. The chapter ends with a discussion of the results. Appendix D provides rational analyses of both recognition phenomena as well as the ABABr (Jung, 1962; Barnes & Underwood, 1959) and crossed associates (Spence & Lair, 1965; McCullers, 1965) cued recall paradigms (discussed in section 2.3).

2.1 Word Frequency in Recognition

The frequency with which a word occurs in general usage is an environmental variable that has been shown to have an affect on recognition performance. Many studies have demonstrated that low frequency words are recognised more accurately than high frequency words (e.g. Gorman, 1961; Schulman & Lovelace, 1970). On the surface, this finding seems to contradict an account based on optimisation to the environment. If high frequency words are seen more often, an optimising mechanism should be adapted to process them more accurately than low frequency words. The word frequency effect, then, represents an important test of the environmental approach.

2.1.1 Experimental Results and Current Accounts

In the typical episodic recognition paradigm subjects are presented with a list of high and/or low frequency words to study. Subsequently, they are asked whether

the words in a second list of high and low frequency words occurred in the first list. Performance is worse for both the high frequency targets and the high frequency distractors (known as the *mirror effect,* Glanzer & Bowles, 1976). Similarly, if subjects are given a forced choice recognition test, that is, they are presented with two alternatives and asked to indicate which was on the list, they perform better on the low frequency words (Glanzer & Bowles, 1976). In addition, the forced choice design allows the comparison of high and low frequency targets, and high and low frequency distractors. Typically, high frequency distractors are chosen over low frequency distractors, while low frequency targets are chosen over high frequency targets (Glanzer & Bowles, 1976).

What makes the low frequency advantage in recognition surprising is that on a number of related phenomena subjects perform better on high frequency words. If the subject is asked to determine if a given letter string is a word (i.e. *lexical decision*), they do so more quickly on high frequency words (Howes & Solomon, 1951). In addition, subjects recall high frequency words more accurately than low frequency words (Gregg, 1976; Deese, 1960; Hall, 1954; Sumby, 1963), although, the pattern is not so clear in mixed lists (see Gillund & Shiffrin, 1984, for a review). Even within recognition, the word frequency effect reverses for very low frequency words with which subjects are unlikely to be familiar (Mandler, Goodman, & Wilkes-Gibbs, 1982; Rao, 1983; Schulman, 1976; Zechmeister, Curt, & Sebastian, 1978).

Many of the earlier accounts of the word frequency effect relied on variables correlated with frequency such as number of associations or concreteness (Allen & Garton, 1968; McCormack & Swenson, 1972; Schulman, 1967), degree of proactive inhibition (Gorman, 1961; McCormack & Swenson, 1972; Shepard, 1967), structural or orthographic distinction (McCormack & Swenson, 1972; Schulman, 1967; Zechmeister, 1969, 1972) and pre-experimental recency (Kinsbourne & George, 1974). Subsequent studies discounted the importance of concreteness (Gorman, 1961) and the number of associates (Kinsbourne & George, 1974), and it has been argued that while the manipulation of these correlated variables does alter performance, frequency has an affect that is independent of any of the variables considered thus far (Underwood & Freund, 1970; Underwood, 1972; Kinsbourne & George, 1974).

The first mathematical model to be applied to the word frequency effect in

recognition was based on the assumption that the recognition decision was made by considering the relative increment in familiarity (Mandler, 1980). If d is the absolute increment caused by studying a word and, H and L are the high and low frequency base familiarity rates respectively, then low frequency words will be better recognised since $\frac{d}{d+L} > \frac{d}{d+H}$. Furthermore, since high frequency words are familiar, they may not be attended to as thoroughly as low frequency words. Consequently, d may be smaller for high frequency words, thus accentuating the difference (Mandler, 1980).

As mentioned above, one of the striking features of the empirical data is that word frequency alters recognition and recall performance in opposite directions. The SAM model (Gillund & Shiffrin, 1984) not only accounted for the low frequency advantage in recognition, but also the high frequency advantage in recall and the diminishing of this effect in mixed lists (i.e. lists that contain high and low frequency words). It was assumed that: (1) the residual association between any two items not rehearsed together is stronger for high frequency words; (2) the association between items that are rehearsed together is stronger for high frequency words; and (3) two criteria, one for high frequency words and one for low frequency words, are chosen. All three of these assumptions depend directly on frequency information.

While the SAM model accounted for the mirror effect, that is, the simultaneous rise in false alarms and decrease in hit rates when low frequency words are compared against high frequency words, it did so by positing independent criteria for high and low frequency words which the subject must adjust appropriately. It has been suggested that an account that dispenses with notions of strength and employs an attention/likelihood model provides a more direct explanation (Glanzer & Adams, 1990; Glanzer, Adams, & Iverson, 1991). Conditions that require greater attention (e.g. low frequency) lead to a more thorough "marking" of the feature vector. In addition, distractor items are assumed to have a background "noise" marking. At test the subject determines the likelihood of the presented item being in the new and old marking distributions and chooses the largest of these.

All of these accounts suggest that additional learning or familiarity should make a given set of words behave more like high frequency words. Allen and Garton (1968), however, conducted a study in which they tested both physics students and

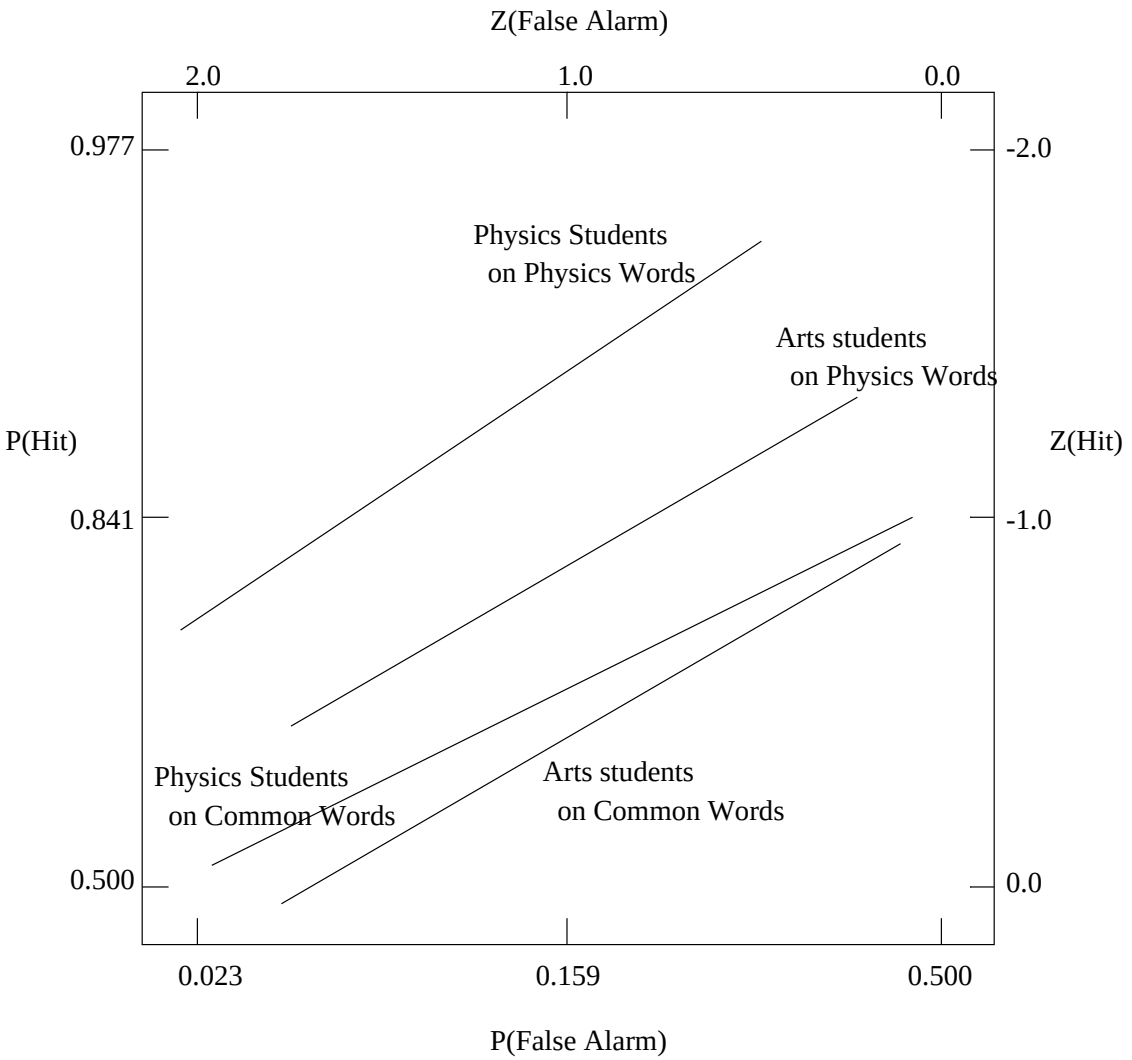

Figure 2.1: Operating characteristics for physics and arts students on physics and common words. Note $P(x) = \int_{Z(x)}^{\infty} N(0,1)dx$ where $N(0,1)$ is the normal distribution with mean zero and standard deviation one. Reconstructed from Allen and Garton (1968).

arts students on common and physics related (i.e. rare) terms. Figure 2.1 shows the operating characteristics they observed. The first point to note is that the comparatively rarer physics words were better recognised than the common words by both student populations. Hence, the usual word frequency effect was found. However, the performance of the physics students on the physics terms was much better than the arts students. Zechmeister et al. (1978), when interpreting these results, suggested that the inferior performance of the arts students indicated that the physics words were acting like very low frequency or nonwords for the arts subjects. However, since the arts students performed significantly better on the physics words than the common words it would seem that they were capable of identifying the words, and, hence, it seems unlikely that the arts students were treating them as very low frequency words. Presumably, the physics students had been exposed to these words with greater frequency than the arts students. To the physics students, then, the physics terms should have appeared to be high frequency and, in keeping with frequency based explanations, they should have performed worse on these items.

The same logic applies to most variables that correlate with frequency. The physics students would be expected to have formed more associations with the physics words and, hence, association explanations predict a decrement in performance (Allen & Garton, 1968). Likewise, if proactive interference from pre-experimental words were the explanation, the physics students who are likely to have been exposed to many similar physics words would be more likely to experience such interference and should have performed worse. The orthographic or structural distinctiveness of the physics words would, if anything, be judged to be weaker by the physics students, since the physics words were familiar. Finally, the pre-experimental recency could not account for the performance of the physics students since it is more likely that they would have seen the words more recently than the arts students. The affect of subject knowledge provides strong evidence that none of the variables examined so far in the literature is the primary cause of the word frequency effect in recognition.

Allen and Garton's (1968) result seems even more striking when compared to a similar experiment in which lexical decision performance was observed for engineers,

nurses and law students on a variety of domain specific terms (Gardner, Rothkopf, Lapan, & Lafferty, 1987). Just as was observed in the recognition experiment, performance was best for subjects who were familiar with the materials. In lexical decision, such a result is not so surprising since high frequency words are identified as words more quickly than low frequency words. These phenomena establish, however, that training seems to uniformly improve performance. Why, then, is the affect of frequency in recognition the opposite to that in lexical decision?

A major difference between the recognition and lexical decision tasks is that the recognition as assessed in the laboratory is episodic, whereas lexical decision is not. Consequently, in the following environmental analysis, statistics relevant to recognition are collected within an episode or context.

2.1.2 Environmental Analysis

When the frequency of a word is calculated, a large number of sources, over a variety of topics and styles, are typically considered. For example, the Francis and Kucera (1982) word frequency statistics, which will be used in the analysis, were collected over a very large sample:

> "The Brown corpus, which was compiled with the view of making it broadly representative of current edited American English, contains selections from five hundred samples belonging to fifteen different genres of writing. The genres range from newspaper reportage to technical writing, and from philosophical essays to various kinds of fiction."

(p 1, Francis & Kucera, 1982)

In contrast, recognition experiments usually occur within a single context. The recognition question is episodic. Subjects are required to recognise an item from a specified context.

Furthermore, Francis and Kucera (1982) note that the distribution of words in different genres is not even. They point out that unlike high frequency words, low frequency words tend to occur in a smaller number of genres. That is, they seem to be context specific, a notion that has some intuitive appeal. For instance, in

a message posted to the connectionist mailing list the word "RAAM"[1] was used six times within an article of some 150 words. The article was about "RAAMs" and, hence, it is not unusual that the word was repeated several times. The word "RAAM", however, is clearly a very low frequency word in English and certainly does not occur in the Francis and Kucera (1982) word frequency inventory.

However, if low frequency words are more likely to be context specific, they may be repeated frequently within the contexts to which they pertain, and hence, the memory system would be optimised to recognise them when they recur within these contexts. The hypothesis to be tested in this analysis, then, is that low frequency words recur more often within a context than high frequency words.

As in the analysis of Anderson and Schooler (1991, i.e. the *New York Times* and email analyses), reading is assumed to require retrieval of the words. More specifically, I presume that as people read text they make an episodic recognition decision for each word and that reading is one of the dominant real world tasks to which episodic recognition performance would be optimised. However, this assumption requires some justification.

Episodic Recognition in Reading: In the case of proper nouns (e.g. John, Mary), an episodic recognition decision must be made. If a person reads a second occurrence of "John", they must recognise that the name has already occurred in this context to realize that the referent is probably the same in each case. A similar argument can be made for pronouns, although, there must also be an episodic cued recall component (i.e. determining to whom the pronoun refers).

Furthermore, the processing of nouns in general would seem to include episodic recognition as a component. For instance, consider the sentence "Shut the door.". If a door has been mentioned in the current context then it is likely that it is the door to which the sentence refers. Realizing that "door" has already occurred is an episodic recognition decision.

While the case is clearest for nouns, other word classes would require episodic recognition also. For instance, if one talks about something that is "green", realizing that green things have already been mentioned in the current episode is important.

[1] RAAM stands for Recursive Auto-Associative Memory a device introduced by Pollack (1988).

Hence, the processing of adjectives would sometimes require an episodic recognition decision. Similarly, verbs may also require an episodic recognition decision. If the word "whistle" recurs there will often be a link between the occurrences.

The major exceptions are the functor words such as "of", "the" or "for". These words recur often and their episodic identity seems to be of little consequence. There is some indication that function words are not processed in a serial fashion and it may be that no episodic recognition decision is made in this case. Further, function words are not usually included in laboratory paradigms. Hence, these words were not included in the analysis (c.f. Anderson & Schooler, 1991).

It might be argued, however, that episodic recognition is not involved in the reading of words because the words from the current context would be transferred to some shorter term data structure[2]. Rather than use a contextual cue to retrieve words from episodic memory such a system would search the short term structure to determine if a word had occurred in the current context.

However, Brannelly, Tehan, and Humphreys (1989) present evidence which suggests that even in a situation in which words might be expected to be in a primary memory structure there would seem to be a retrieval component which is subject to proactive interference and, hence, would seem to involve episodic (secondary) memory. In their second experiment, Brannelly et al. (1989) presented subjects with memory sets of either two or four items. After a filled 14 second delay, subjects were presented with a probe which they were required to identify as having been in the memory set or not. A second probe was then presented and again the subjects were required to determine if it had been present in the memory set. In addition, proactive interference was manipulated by choosing items from the same taxonomic category on three consecutive trials. The assumption was that subsequent to the presentation of the first probe the memory set would be retrieved from secondary memory into primary memory and scanned to establish the presence of the probe item. When the second probe was presented one would expect that the memory set would still be present in primary memory and, hence, should not be affected by proactive interference which is usually a feature of retrieval from secondary memory. However, there was a significant amount of proactive interference. Hence, even if

[2]I would like to thank an anonymous reviewer for suggesting this possibility.

the words from the current context are present in some short term data structure it is likely that there will also be retrieval from secondary memory.

It could be argued, however, that the short term data structure would be longer term than the primary memory which was was implicated in the memory probe tasks used by Brannelly et al. (1989) and may be subject to proactive interference. The question becomes "Can such a memory system be distinguished empirically from episodic memory?".

Humphreys and Burt (1994) argue on the basis of data from Jacoby (1983, , experiment two), that episodic information affects repetition priming even when context is not reinstated, but is implicit because the test occurs immediately following the study opportunity. In Jacoby's second experiment, subjects were presented with a study list and then with a surprise perceptual identification task. The saliency of the study items was manipulated by administering the test either immediately in the high saliency condition, or after 24 hours in the low saliency condition. In both cases the test list consisted of 10 words from the study list and 90 new words. The subjects were not informed about the relationship between the study and test lists. The perceptual identification of old words was facilitated in the high saliency condition relative to the low saliency condition. Jacoby (1983) also found inhibition of the new words in the high saliency condition relative to the low saliency condition. As Humphreys and Burt (1994) point out, such a finding can be explained by assuming that an episodic memory system is employed in both conditions and that the context cue in the high saliency is more similar to that at study than the context cue from the low saliency condition. It is also possible to explain these results using a short term data structure by supposing that the subjects were using this structure to make their decisions in the high saliency condition and were not in the low saliency condition. Under such a view, the old words would show facilitation in the high saliency case because the short term store is searched first. The inhibition of the new words in the high saliency case relative to the low saliency case would occur because of the lose during the search of the short term datastructure that occurs in the high saliency case but not in the low saliency case.

What this analysis suggests is that it may be difficult to distinguish the consequences of the short term data structure from an episodic memory system which

allows the use of context cues which persist as a consequence of continuity rather than requiring explicit reinstatement.

In terms of using reading as the environmental analog of episodic recognition tasks, the distinction may not be important. If there is a short term structure, one must assume that it is under the strategic control of the subject to account for the Jacoby results. If this is the case, it would also be available to subjects in making recognition judgements in the "episodic" tasks that have been studied. Regardless of whether it is a list context that is reinstated or a short term datastructure that is identified as relevant it is still the recurrence probability that one would expect to be predictive of performance on these tasks.

The above discussion is not meant to suggest that episodic recognition is the only subtask of reading, nor that reading is the only task in which episodic recognition takes place. As noted above, reading must include cued recall in order to determine to whom or what pronouns refer, and there are most probably many other situations in which cued recall would be implicated. Other tasks, notably conversation, would also seem to require episodic recognition. Reading has been chosen, in this case, because it is a very common task for which large corpora are readily available for analysis.

Method

Three hundred messages from the connectionist mailing list were used to examine the recurrence probabilities of high and low frequency words. Each message was considered as a separate context. A more appropriate way of separating the list may have been into *threads* (i.e. series of messages about the same topic), however, it was deemed more valuable to maintain the objectivity of the context boundary placement. In addition, all of the messages occur in the fairly specific context of the connectionist mailing list. Words such as "neural" or "connectionist" are likely to occur across message boundaries within the list but would have low probability of occurring across message boundaries in a newspaper, for instance. Both of these factors serve to diminish the affect of frequency on recurrence probability.

Before counting began, the messages were filtered to remove mail headers, ftp instructions, punctuation, misspellings, and function words. Upon completion of

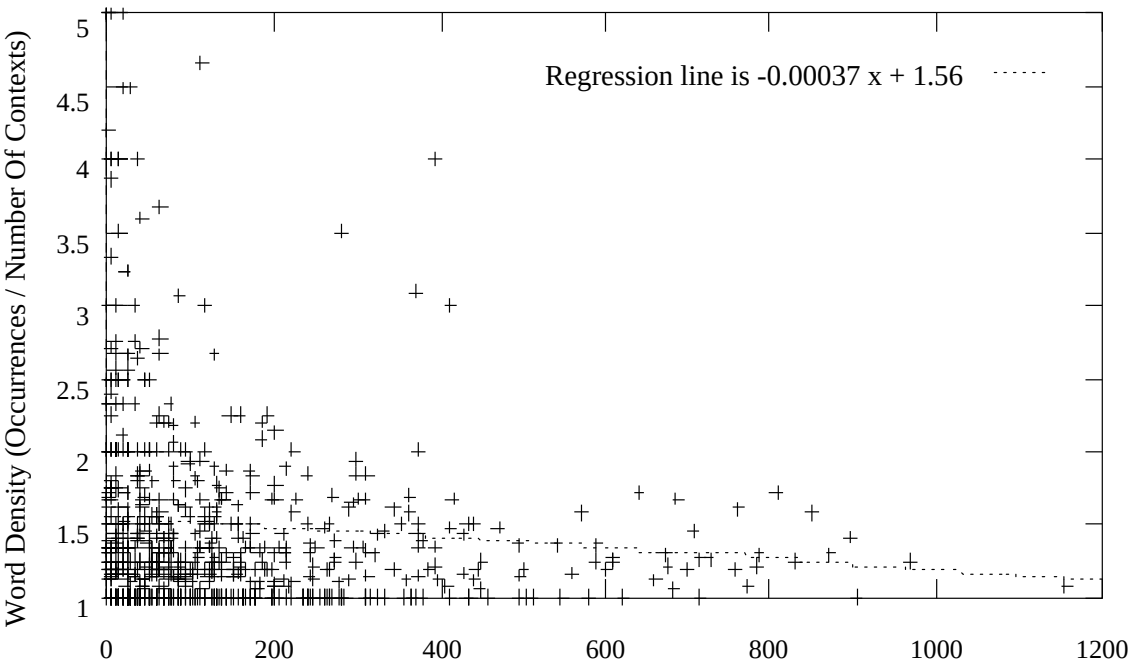

Figure 2.2: Word density versus word frequency. The word density, defined as the average number of occurrences of a word per context in which it occurred, is plotted against the word frequency taken from Francis & Kucera (1982). The regression line is negative indicating that the less frequent words are denser, that is, they are more likely to recur in the same context.

this process 758 unique words remained.

To quantify the likelihood of recurrence of a word within the same context, the *word density* was defined as the number of occurrences of the word divided by the number of contexts in which it occurred.

Results

Figure 2.2 is a scatter graph of the word density plotted against word frequency values taken from Francis and Kucera (1982). The slope of the regression line is negative, $F(1,758) = 9.186$, $p < 0.0025$, indicating that low frequency words are more likely to recur within a context. To give an indication of how word density varies as a function of word frequency, the *mean* word density was plotted against word frequency (see figure 2.3). When plotted on a log-log scale, the graph is nonlinear suggesting a greater than polynomial rate of decline. However, the high frequency points are near the minimum word density of one and the form of the function may be influenced by floor effects making interpretation difficult.

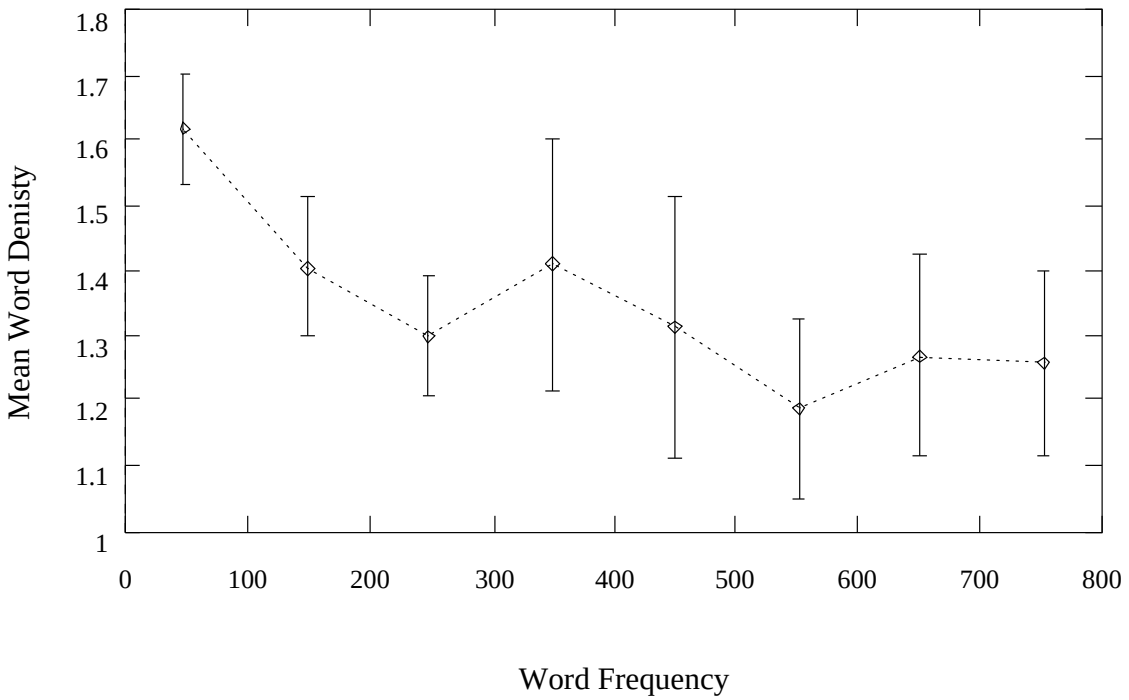

Figure 2.3: Mean word density verses word frequency. The bars represent the 95% confidence intervals.

Discussion

For the connectionist mailing list corpus examined in this study, low frequency words recur more often than high frequency words within a context. To give an indication of the magnitude of the result, consider the word "access", which occurs 24 times per million in the Brown corpus (Francis & Kucera, 1982). Its word density in the connectionist mailing list corpus was 1.556, which is very close to the value that would be predicted on the basis of the regression line (i.e. 1.559). In contrast, the word "made", which occurs 1156 times per million in the Brown corpus, has a word density of 1.077 (1.132 predicted on the basis of the regression line). If one subtracts the baseline (i.e., 1.0), the word density of "access" is over seven times the word density of "made". Hence, the effect is not only significant but large as well.

The current data support the recurrence rate explanation of the word frequency effect. Since low frequency words are more likely to recur, an optimising memory system should adapt to perform more accurate episodic recognition on the low frequency words. However, by definition, the occurrence rate of a high frequency word that has not yet occurred in the current context will be greater than that of a low frequency word. Given the above analysis, there must be an interaction between

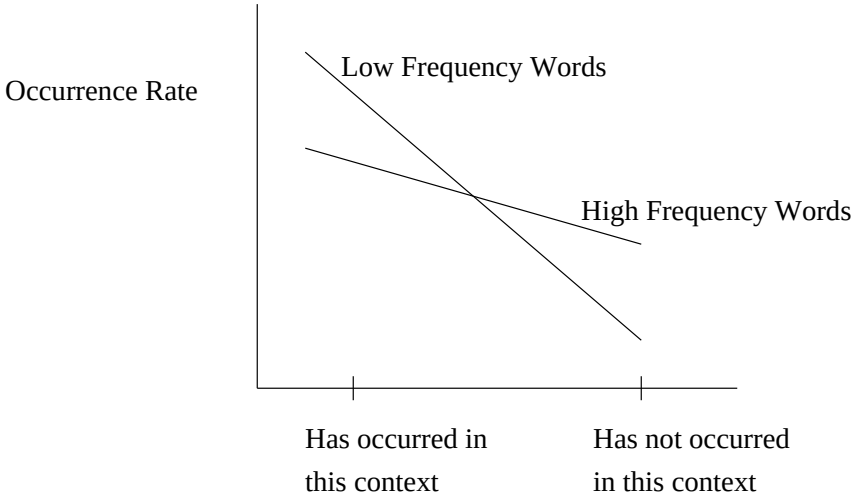

Figure 2.4: An idealised graph of the word frequency by occurrence history interaction. The environmental data presented demonstrate that low frequency words have a higher recurrence probability than high frequency words. By definition high frequency words occur with greater frequency, however. Hence, if the word has not yet occurred the probability of needing a high frequency word will be greater.

occurrence history and word frequency (see figure 2.4).

Furthermore, if one considers the probability of recurrence at any given instant as a function of the time since the last occurrence of the word, low frequency words must decay more rapidly than high frequency words (see figure 2.5).

The interaction of word frequency and occurrence history is particularly interesting in the light of the work by Glanzer and Bowles (1976). They conducted forced choice experiments including conditions in which subjects were required to choose between a high frequency and a low frequency item both of which had been on a study list, and a high frequency and a low frequency item neither of which had been on the study list. They found that when both items had been seen, subjects were more likely to choose the low frequency item. When neither of the items had been seen, subjects were more likely to choose the high frequency item. If one assumes that performance is proportional to probability of recurrence, the Glanzer and Bowles' result maps directly onto figure 2.4.

Recurrence rate also provides an explanation of the affect of word knowledge (Allen & Garton, 1968). While low frequency words have a higher recurrence rate in general, the recurrence rate of the physics words for physics students is likely to be greater still. The memory systems of the physics students would be optimised to the physics words and, therefore, it would be expected that they should perform

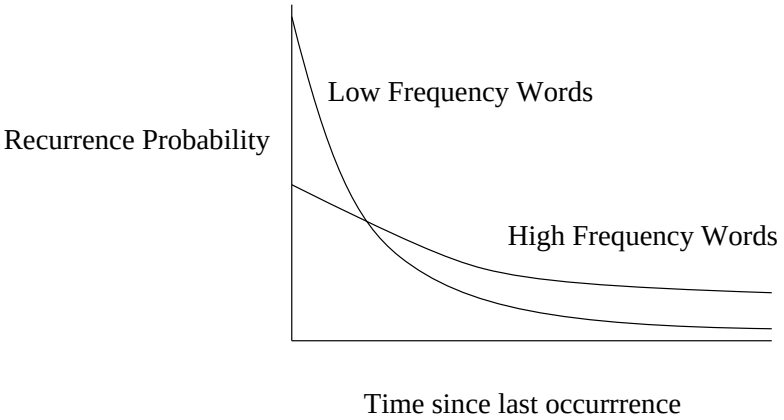

Figure 2.5: Recurrence probability of a word verses time since its last occurrence. Low frequency words have a greater recurrence probability than high frequency immediately after their first occurrence. Given that high frequency words asymptote to a greater recurrence probability, the decay rate of low frequency words must be greater than that of high frequency words.

better on these words than the arts students.

In addition, recurrence rate provides a straightforward account for why lexical decision should show the reverse effect to episodic recognition. When reading text a lexical decision would be made at each occurrence of a word. In contrast, a positive episodic recognition decision can only be made at the recurrence of a word in a context. Hence, lexical decision should favour high frequency words, which have a higher occurrence rate, while episodic recognition should favour low frequency words, which have a higher recurrence rate.

There is, however, a caveat to the above discussion. While a given subject may have learned that particular low frequency words tend to recur often, it is questionable whether they could have been exposed to a sufficiently large subset of the low frequency words to extract the recurrence information on a case by case basis. While the above reasoning would suggest that occurrence and recurrence frequency are the prime determinants of the frequency effect, other variables related to input representation (e.g. orthographic distinctiveness) or functional role (e.g. concreteness, or number of associates) may contribute to the establishment of a similarity landscape in which the representations of the low frequency words are separated from those of the high frequency words[3]. Rather than learn the recurrence

[3]Such a similarity landscape, in the context of a distributed model of memory, could also provide the basis of an explanation of the release from proactive interference that is observed when a frequency category is changed (Swanson & Wickens, 1970).

information for each word individually, an optimising mechanism might determine that the class of low frequency words are more likely to recur. Nonwords and very low frequency words may not show the same performance pattern as low frequency words (Zechmeister et al., 1978) because they are not sufficiently similar to the low frequency words for which the recurrence response has been learned.

The above environmental analysis together with an optimising account of memory provide a parsimonious explanation of the affect of word frequency in recognition.

2.2 The Null List Strength Effect in Recognition

It is perhaps not surprising that environmental analysis proved useful in the explanation of the word frequency effect. Word frequency is an environmental variable. But, can environmental analysis shed light on phenomena that are not so obviously related to the statistics of the environment? To explore this question this section examines the null list strength effect in recognition (Ratcliff et al., 1990; Shiffrin et al., 1990).

The presence of repeated items in a list to be recognised does not affect performance on non-repeated items. This phenomenon, know as the *null list strength effect*, has undergone substantial research both empirical (Murnane & Shiffrin, 1991a; Ratcliff et al., 1990) and model theoretic (Shiffrin et al., 1990; Murdock & Kahana, 1992; Chappell, 1993; Chappell & Humphreys, in press, 1993; Heathcote, 1993) in recent years. Unlike the word frequency effect studied in the previous section, however, the list strength effect is not so obviously related to the statistics of the environment. The purpose of this study is to investigate what impact the environment might have on the phenomenon.

2.2.1 Experimental Results and Current Accounts

An issue that has been of interest for more than a century is the affect on memory of study lists that contain items of different strengths (Ebbinghaus, 1964; Thorndike, 1965; Tulving & Hastie, 1972; Ratcliff et al., 1990). The strength of an item is manipulated either by increasing the number of times it occurs in the study list or by increasing the length of time for which it is studied. Items of greater strength are

Table 2.1: Accuracy results for nonrepeated items in different list types.

Test Type	AB vs ABB	AB vs ABC	ABB vs ABC
Recognition	$=_{(<)}{}^{*}$	$>$	$>^{*}$
Cued Recall	$=_{(>)}{}^{*}$	$>$	$>^{*}$
Free Recall	$>$	$>$	$<$

Note. Asterisks indicate the results that were not predicted by the mathematical memory models. $=_{(<)}$ means equal or perhaps slightly worse. $=_{(>)}$ means equal or perhaps slightly better. AB is a short list without repeated items. ABB is a long list with repeated items. ABC is a long list without repeated items. This table is compiled from results by Tulving & Hastie (1972) and Ratcliff, Clark & Shiffrin (1990).

generally retrieved more accurately than weak items (Ebbinghaus, 1964; Thorndike, 1965). Of more recent interest is what happens to performance on the weak items as a consequence of strengthening other items within the list. Table 2.2.1 outlines the pattern of accuracy results for the weak items within different list types for recognition, cued recall and free recall.

For free recall, Tulving and Hastie (1972) demonstrated that performance on unstrengthened items dropped significantly in comparison to the AB and ABC control conditions. For recognition, however, strengthening an item does not degrade performance significantly on the non-strengthened items (known as the *null list strength effect,* Shiffrin et al., 1990; Ratcliff, 1990; Yonelinas et al., 1992). This result contradicted the major mathematical memory models all of which predicted that performance on A in the ABB condition should be as poor if not worse than that in the ABC condition (Ratcliff et al., 1990).

Since the initial experimental work on the null list strength effect, a number of researchers have attempted to alter their models to account for the data. Shiffrin et al. (1990) made two modifications to the SAM model. The first was to assert that when an item was strengthened, either by additional study time or additional study presentations, the same representation (or image) in memory was affected. Secondly, strengthening an item not only increased its context-to-item association, but also decreased its similarity to all other items. This process is called *differentiation* and is illustrated in figure 2.6. In the figure, familiarity is measured by a dot product. Strengthening an item has the effect of both increasing the length of the vector (i.e. context-to-item association) and making it closer to orthogonal to the other item

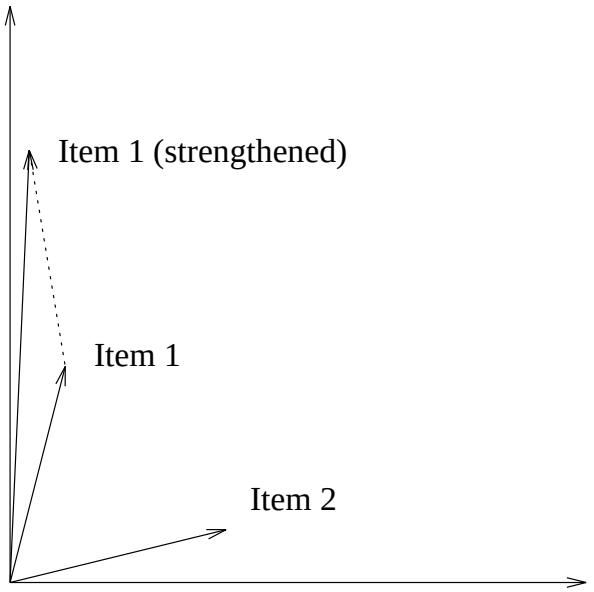

Figure 2.6: Differentiation in the SAM model. Strengthening an item increases its context-to-item associations, but decreases its similarity to other items. In the diagram the context-to-item association is indicated by the vector length, while similarity is represented by the angle between the vectors. When *item two* is used as a cue, the familiarity (i.e. dot product) can be made identical in the strengthened and unstrengthened cases by appropriate choice of the parameters.

vector. By setting the parameters appropriately, the strengthening of *item one* can be made to have little affect on the familiarity calculated when cuing for the weak *item two*. Even though the length of *item one* has increased, it has also become less similar and, hence, the familiarity remains constant. If the model is cued with *item one* the additional length of the strengthened vector can still produce the main effect of strength. Likewise, increasing the number of study vectors will lead to greater interference and a length effect will be observed.

The model by Chappell and Humphreys (in press, 1993), while implemented in a connectionist architecture based on the matrix model (Pike, 1984), relies in part on a similar principle to SAM when explaining the null list strength effect. When an item is strengthened, context-to-central weights are updated (probabilistically). While the representations in the Chappell and Humphreys (in press, 1993) model are distributed rather than local, the affect is to increase the context-to-item association. At the same time, the weights of a central autoassociator are increased and a global inhibition is learned. The net affect of these manipulations is to make strengthened items more differentiated. Hence, the model accounts for the null list strength effect and demonstrates the difficulty in distinguishing between distributed

and nondistributed models of human memory[4].

One way of eliminating the list strength effect is to assume that pre-experimental items are also present in memory. This approach was used by Murdock and Kahana (1992) in their extension of the TODAM model. The affect of adding additional items is to swamp with noise any difference between the mean familiarity of an item that occurred with unstrengthened items and an item that occurred with strengthened items. While such a method can decrease the list strength effect to negligible levels, it also undermines the prediction of the list length effect and, hence, is not sufficient to explain the pattern of results (Chappell, 1993). This example illustrates the importance of modelling the entire pattern of recognition results. It is the building of a model that captures the list strength effect while also accounting for the item strength and list length effects that has proven difficult.

All current accounts of the null list strength effect presume that it is the mechanism of memory that is primarily responsible for the phenomenon. As yet, there has been no empirical characterisation of the environmental context of the effect. The next section aims to characterise the environmental statistics relevant to the item strength, list length and list strength effects. Some of this work has already been done by Anderson and Schooler (1991) and, hence, the analysis starts with a review of their work. Then, an environmental analysis that addresses the list strength effect and re-addresses the item strength condition is conducted.

2.2.2 Environmental Analysis

Review of Previous Environmental Analysis

Anderson and Schooler (1991) examined a number of sources (*New York Times*, parental speech, electronic mail) to determine the probability that an item will occur given its frequency, recency and pattern of prior exposures. The first two results will be used in this thesis and will now be summarised.

The probability that an item will recur is linearly related to the frequency of the item (Anderson & Schooler, 1991). This relationship was found in all three sources.

[4]Heathcote (1993) has also used a connectionist model called contextual ART (see Carpenter & Grossberg, 1988) to account for the list strength effect. However, the complete details have yet to be published.

In the *New York Times* data the constant of proportionality was 1.0, in the parental speech data it was 0.76 and for the electronic mail data it was 0.9 (see figure 2.7). Such a result is to be expected since words occur with a certain frequency and barring some other form of intervention, it is to be expected that they will continue to occur with the same frequency. The constants of proportionality were less than one because new words were entering the sample (Anderson & Schooler, 1991), particularly in the case of parental speech.

In addition to frequency information, Anderson and Schooler (1991) studied the affect of retention interval. In this case, they plotted the probability of an item recurring as a function of the time since its last occurrence (see figure 2.8). They found that recurrence probability showed a similar negatively accelerating curve as seen in the laboratory when interval is manipulated. Furthermore, when the log need odds were plotted against log duration, a linear relationship emerged indicating that the original curve was a power function. In the *New York Times* data the exponent was 0.73, in the parental speech data it was 0.77 and in the electronic mail data it was 0.83.

Anderson and Schooler (1991) provide many of the statistics required to characterise the affect of repetition in the environment. However, there are still two issues that must be resolved to give a complete picture. The first involves the independent affect of contextual frequency. Anderson and Schooler (1991) did not control for the affect of general word frequency. In the laboratory, however, it is the frequency within the study context that is manipulated. The linear relationship between occurrence frequency and recurrence rate, which Anderson and Schooler (1991) found, may have been a consequence of the general word frequency of the item. In the following study, the general word frequency has been factored out so that the affect of frequency in the current context can be assessed.

Secondly, the following study will address the question of how the recurrence probability of a word is affected by the occurrence probabilities of other words in the context - the environmental analogue of the list strength effect.

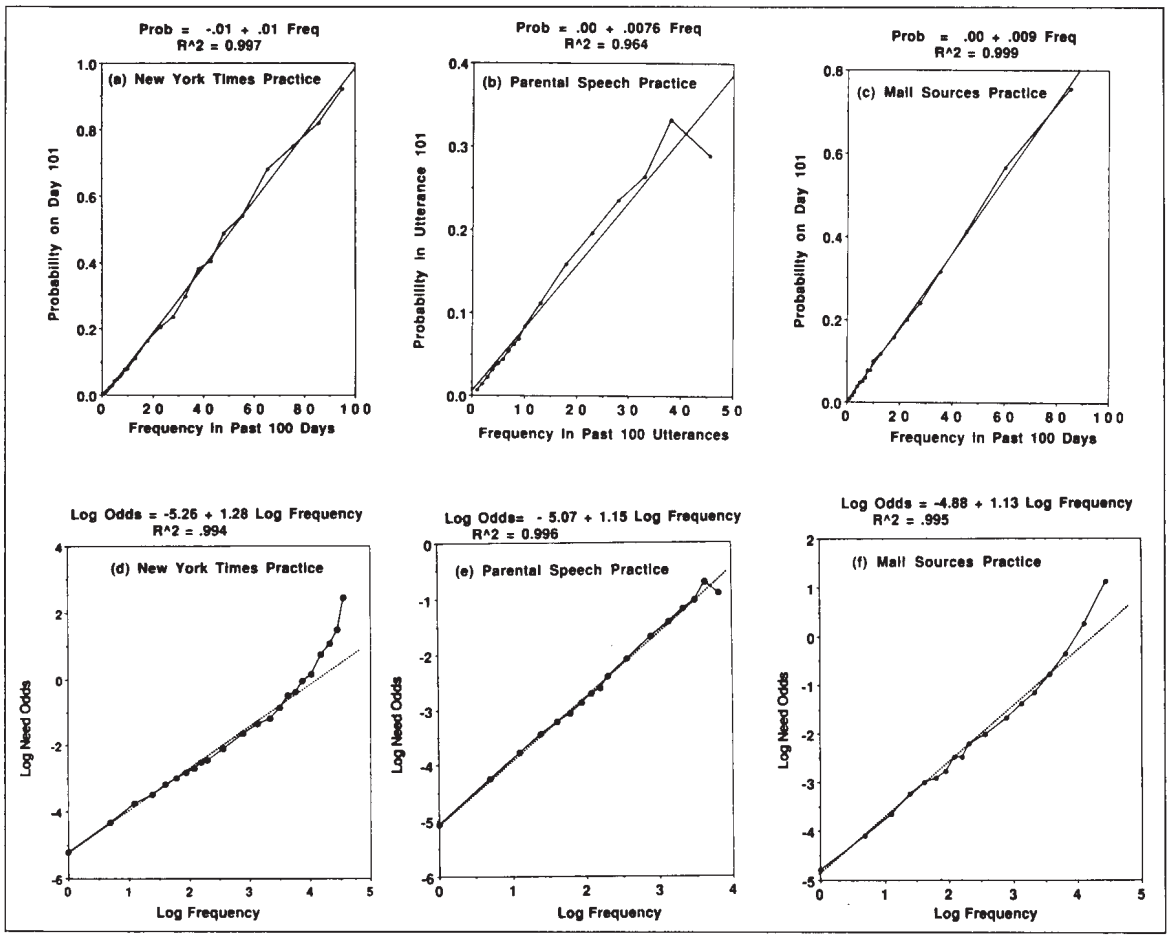

Figure 2.7: Recurrence probability as a function of frequency. (a) Probability of a word occurring in a headline of the *New York Times* on Day 101 as a function of the number of times it occurred in the previous 100 days; (b) probability of a word occurring in the 101st utterance from a parent as a function of the number of times it occurred in the previous 100 utterances; (c) probability of receiving a message on the 101st day from a source as a function of the number of times messages were received from that source in the previous 100 days. Panels (d-f) provide transformation of (a-c) plotting log needs against log frequency. Reproduced from Anderson & Scholar (1991), figure six.

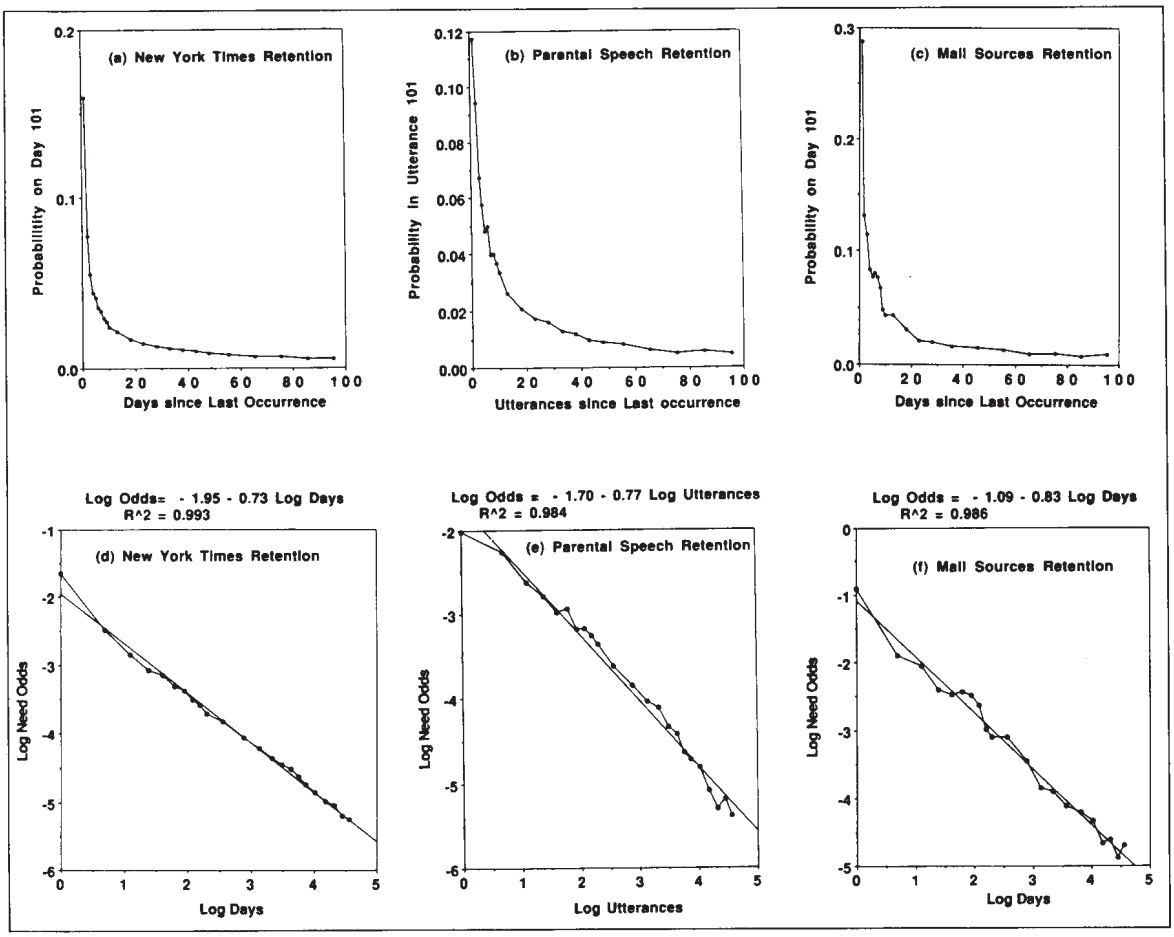

Figure 2.8: Recurrence probability as a function of recency. (a) Probability of a word occurring in a headline in the *New York Times* on day 101 as a function of how long it has been since the word previously occurred; (b) probability of word occurring in the 101st utterance from a parent as a function of how many utterances it has been since the word previously occurred; (c) probability of receiving a mail message from a source as a function of how many days it has been since a message was last received from that source. Panels (d-f) provide transformation of (a-c) plotting log need odds against log frequency. Reproduced from Anderson and Scholar (1991) figure seven.

Analysis of the Item and List Strength Effects

To determine whether the environmental statistics predict a list strength effect it is necessary to establish if:

$$\begin{aligned} &P(W_i \in TestSegment \mid \\ &\quad W_i \in StudySegment \,\&\, W_{j\neq i}\, occur\ frequently\ in\ study\ segment) \approx \\ &P(W_i \in TestSegment \mid \\ &\quad W_i \in StudySegment \,\&\, W_{j\neq i}\, occur\ infrequently\ in\ study\ segment) \end{aligned} \tag{2.1}$$

Hence, the first step in the analysis was to characterise the strength history of a given item. The number of repetitions of the word was chosen as an indication of strength rather than the amount of time the reader spent processing the word because in the environmental setting it is difficult to control or estimate the amount of time an individual spends on each item. In the two experimental lists (i.e. AB and ABB), the number of *unique* items is constant while the total number of items increases in the ABB list. Therefore, the general degree of repetition, $\frac{Number\ of\ Items}{Number\ of\ Unique\ Items}$, is greater in the ABB list than in the AB list. The sample text can be divided into segments that have equal numbers of unique items. The number of actual items will vary, however. Some segments will have large numbers of actual items and will parallel the ABB condition. In other segments, the amount of repetition will be lower, parallelling the AB condition.

The list strength effect occurs in the experimental setting if increasing the general repetition rate decreases performance on the nonrepeated items. Correspondingly, if general repetition rate decreases the recurrence rate a list strength effect will be observed in the environmental setting.

In addition, it is important to establish that the main effect, that is, that repetition improves an item's performance, is evident in the environmental statistics. That is,

$$\begin{aligned} &P(W_i \in TestSegment \mid W_i\, occurs\ frequently\ in\ study\ segment) \;> \\ &P(W_i \in TestSegment \mid W_i\, occurs\ infrequently\ in\ study\ segment) \end{aligned} \tag{2.2}$$

If there is no affect on non-repeated items, as the experimental evidence on the

list strength effect suggests, the study would provide only very weak evidence unless the main effect, in which a positive effect is expected, is also tested. Hence, the *target repetition rate*, that is, the number of times the target item occurred divided by the length of the segment, was included as a variable in the analysis. In the experimental setting, increasing the target repetition rate has been found to improve performance and, hence, increasing the target repetition rate is expected to increase recurrence probability.

In the uncontrolled circumstance of general text, however, the target repetition rate may be significant purely as a consequence of general word frequency. Hence, word frequency was included as a variable in the analysis to ensure that the target repetition rate was significant independent of any contribution from general word frequency.

To summarise, on the basis of the experimental result it would be expected that general repetition rate would have no effect of recurrence probability, but that target repetition rate would have a positive effect.

Method

Raw Data: Articles were taken from the online version of the University of Minnesota Daily. This data set was chosen because it is easily and globally accessible[5] and includes a significant range of topics in a journalistic writing style, which was assumed to be representative of the sort of text human subjects commonly read. Appendix B.1 shows an example of the raw input. There were 260 articles (i.e. contexts) used.

Clean Data: Noise components such as punctuation, function words and numerals were deleted. In addition, case was standardised to lower, allowing the counting program to recognise repeated items. The result of applying this process to the raw sample in appendix B.1 can be found in appendix B.2.

Dividing the Text Corpus: In order to mimic the study/test paradigm employed in typical memory experiments, a study segment and a test segment were

[5]The Minnesota Daily articles are available for ftp on internet and can be accessed using the gopher program.

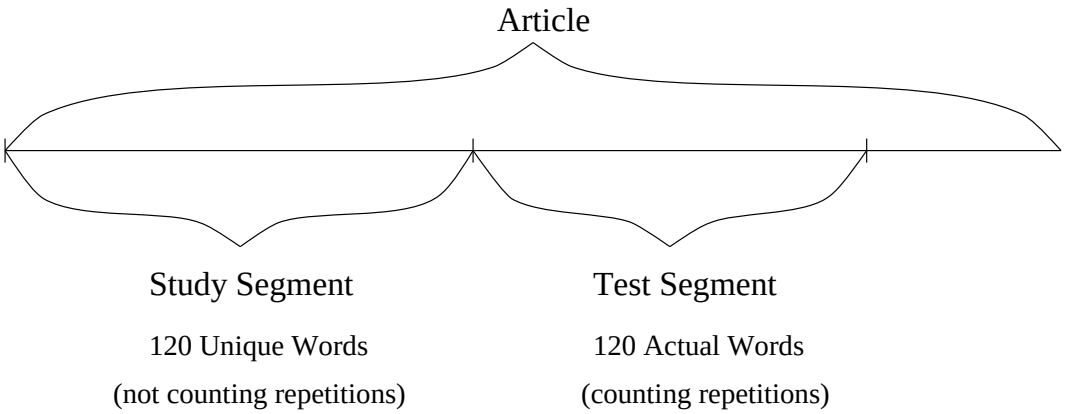

Figure 2.9: The division of articles in the environmental study of list strength. Each article was divided into study and test segments. The study segments contained 120 *unique* words, although the fact that words may be repeated means that it is likely that there will be more actual words in the segment. The test segments were always 120 actual words long.

cut from the start of each message. The study segments contained 120 *unique* words, although the fact that words may be repeated means that it is likely that there will be more than 120 words in total. The test segments were always 120 actual words long. If an article was not long enough to extract these segments it was discarded. Upon completion of this process 198 articles were retained. Figure 2.9 shows diagrammatically the division of an article.

Appendix B.3 shows the study and test segments extracted from the article in appendix B.1. The 120th unique word occurs at the 135th actual word, meaning there are 15 repetitions, with words such as "baseball" and "clock" occurring multiple times.

Segments of size 120 were chosen in response to the size of the articles. Larger segments provided more accurate statistics, but required many articles to be discarded since they would not have been long enough to allow extraction of both segments. A size of 120 provided a balance between these conflicting concerns.

Calculating Recurrence Probability: The dependent variable in the analysis was the *recurrence probability*, a measure of how likely a given word was to recur at a position in the test segment. This statistic was calculated by taking the number of repetitions and dividing by the total number of words in the test segment (which was always 120). The word "baseball" occurs six times in the example test segment and, hence, has a recurrence probability of $6/120 = 0.05$.

Calculating General Repetition Rate: The *general repetition rate* was included as a measure of the repetition within the study segment. It was calculated by dividing the number of actual words by the number of unique words (which is always 120). In the study segment in appendix B.3, the number of actual words is 135, hence, the general repetition rate was $135/120 = 1.125$. The more items that are repeated the higher this measure becomes.

Calculating the Target Repetition Rate: The *target repetition rate* for each word was calculated by taking the number of repetitions of a word within the study segment and dividing by the total number of words in the segment. In the sample of appendix B.3, there are five occurrences of the word "baseball" in the study segment. Since there are 135 actual words, the word "baseball" has a target repetition rate of $5/135 = 0.037$.

Calculating Word Frequency: The word frequency is calculated over all articles and is simply the number of times the word occurs. For example, in the Minnesota sample the word "baseball" occurs 27 times within the 198 articles and, hence, its word frequency would be 27.

Removing Points: If a word occurred in the test set but not in the study segment it was not a possible target and was discarded for the purposes of the analysis. In addition, if the word appeared in the study segment but not in the test segment its recurrence probability was considered to be at floor and it was discarded. Of 17390 initial observations 3716 remained after filtering.

Results

The 198 articles, when cleaned, had an average length of 436 words with a standard deviation of 147. Table 2 shows the means and standard deviations of each of the variables.

Since general repetition rate was defined over study segments whereas target repetition rate and word frequency were defined on individual words there were multiple observations that had the same general repetition rate. If all observations were analysed together the variance of the general repetition rate would be restricted

Table 2
Parameters of the Dependent and Independent Variables

	Mean	Standard Deviation
Recurrence Probability	0.014	0.010
General Repetition Rate	1.304	0.122
Target Repetition Rate	0.012	0.010
Word Frequency	131	259

and it would be less likely to reach significance. To avoid this problem, only one observation from each article was taken and the analysis was run 20 times without replacement.

Despite this precaution, the multiple regression analysis showed no significant affect of general repetition rate on 19 of the 20 trials. The target repetition rate was significant on 17 of 20 trials and the word frequency varied the most showing significance on 14 of 20 trials. Appendix C shows the significance level of adding each of the variables given that the other two variables have already been added to the model for each of the 20 samples.

Discussion

As in the experimental setting, no affect of general repetition rate on recurrence probability was found. The degree of repetition shown by other words had no impact on the likelihood that a given word recurred.

From an environmental perspective this result is not very surprising. Why should the rate of repetition of other items affect the recurrence probability of an item? It is only when one looks from a mechanistic perspective that the lack of a list strength effect seems unusual.

In addition, it was found that the need probability is a monotonically increasing function of the frequency of recent use (Anderson & Schooler, 1991). However, the current study takes the Anderson and Schooler (1991) analysis one step further. A high frequency word may occur frequently in both the study and test segments by virtue of the fact that it is a high frequency word. Hence, a positive relationship between study and test set frequency may have been expected independently of whether the occurrence of a word "primes" the memory system to expect the word

again within the current context. This study shows that the target repetition rate is significant even when the affect of word frequency is factored out. Hence, there is a basis for confidence in the power of the test and the finding that recent repetition is positively correlated with recurrence probability.

In conjunction with the Anderson and Schooler (1991) results, we now have a complete picture of the environmental context of the word frequency and list strength effects in recognition. So far, however, only recognition paradigms have been considered in the thesis. The next section discusses the extension of the analysis to cued recall paradigms.

2.3 Towards an Environmental Analysis of Cued Recall

Episodic cued recall is manifest in many everyday tasks. In fact, any task in which there is a general pattern of recurrence of information that is altered on occasion will qualify. For example, one may develop a memory that binds a time slot to a television program that is broadcast on a regular basis. The memory system might learn, for instance, that Star Trek occurs on Wednesday nights at 7:30pm. On special occasions, such as the Olympics or a football final, these events may be broadcast in preference to the regular schedule and the pattern may be broken. Generally, such breaks are well advertised and so the memory system is likely to be exposed to the fact that, for this week, the program will be different. Note that, the memory system seems to be able to retain the new information despite the fact that the pre-existing memory may be well established. In addition, the previously established pattern is re-assumed the following week, and the fact that the pattern changed last week does not override this memory[6]. For an optimising memory system maintaining episodic information would be important. The recurrence probability for the new information will be high while it is applicable and in the weeks after the event the recurrence probability for the established pattern will be greatest.

Not all cued recall paradigms have clear environmental analogues, however. In

[6]This pattern resembles the ABAD paradigm with the exception that instead of A and B being paired in a first list A and B are pre-existing associates.

particular, environmental analysis of the ABABr (Jung, 1962; Barnes & Underwood, 1959) and crossed associates paradigms (Spence & Lair, 1965; McCullers, 1965), which require three way joint information (Humphreys, Wiles, & Bain, 1992), may prove difficult. These paradigms involve designs in which pre-established pairs (from pre-experimental learning or previous lists) are crossed. For instance, in an ABABr experiment the following lists might be presented:

List 1

Table Road

Grass Violin

List 2

Table Violin

Grass Road

The same words appear in the same positions within a pair. All that has changed is which words are seen together. At test, subjects are given the list number and the first word of the pair and are required to recall the second word of the pair. For example, the subject might be asked which word occurred with "Table" in "List 1". In the above lists, the correct response would be "Road". As Humphreys et al. (1992) pointed out, it is not possible to determine the correct answer from two-way associations. Both "Table" and "List 1" are associated once with "Road" and once with "Violin". On the basis of two way associations subjects would respond with "Violin" as often as they responded with "Road". In practice, however, subjects are able to respond well above chance on this task (Humphreys et al., 1992).

Finding an environmental task that exhibits the crossed design is difficult and it is not clear that such situations occur with sufficient frequency to warrant an optimising memory system taking them into account. The ABABr and crossed associates paradigms have been computationally rather than ecologically motivated.

However, for cued recall in general, there are environmental tasks that can be analysed. In the section on reading and episodic recognition, it was argued that a positive episodic recognition decision was made at each recurrence of a word. In many cases, however, it seemed at least as likely that episodic cued recall would also be involved. For instance, in the case of pronouns, it is necessary to recall the

name of the person to whom the pronoun refers. Similarly, it could be argued that when reading the name "John", subjects recall many details about "John" that only apply in the current context (e.g. the colour of the jumper he is wearing). Similar arguments can be made for other classes of words although functor words, as discussed above, may be exceptions.

Conversation may prove even more useful in the analysis of cued recall. Unlike recognition, cued recall involves the production of an item. While it is possible to speculate as to the recall processes that may be occurring during reading, the person reading text is not compelled to produce items. When in conversation, however, the participants are constantly producing utterances. As an approximation, one could consider one utterance as the cue set for the production of the next speaker's utterance. The statistics to be measured would be the probability of production of a word given the words in the preceding utterance, as a consequence of previous co-occurrence. Conversation, then, provides directly observable evidence of the statistics of cued recall in the environment.

2.4 Discussion and Conclusions

What does the environment look like? The environmental analysis in this chapter provides evidence in support of the following propositions:

1. Within a context, words of low frequency in the language are more likely to recur than words of high frequency in the language.

2. Words that have occurred frequently in the context are more likely to recur than words that have occurred infrequently in the context.

3. The occurrence frequency of other words in the context does not affect the recurrence probability of a word in the context.

While it is too early to make generalisations, and there are paradigms (such as the ABABr and crossed associates) for which environmental analysis is difficult, the success in providing environmental explanations of the word frequency and null list strength effects suggests that there is much to be gained by considering the statistical landscape in which memory models are situated. The value of the approach is best

illustrated by the list strength result. It was found that the degree of repetition of other items within a text segment does not affect the recurrence probability of a word. It could be argued that this result is exactly what one would expect and that there was really little need to go to the extent of testing the environmental statistics. Under a mechanistic framework, however, the list strength effect seems counter intuitive. The issues that seem relevant in the mechanistic framework are the strengths and types of items that will interfere with the target word. Implicit in the mechanistic framework is an assumption about the conservation of "memory resource". If an item is added or strengthened then it takes more memory resource resulting in a corresponding decrement in some other memory. The list strength effect in recognition violates this assumption and consequently all of the major memory models failed to predict the phenomenon.

The word frequency result is a little more unexpected. However, since recurrence rate is most relevant to episodic recognition and occurrence rate is most relevant to lexical decision, the environmental evidence provides insight into why word frequency affects these tasks differently. Furthermore, the affect of word knowledge, which undermines previous explanations of the word frequency effect, is explained naturally by recurrence rate and an adaptive memory system.

Thus far, the nature of this adaptive memory system has not been elaborated. In the next chapter, some criteria for such a system are outlined and the Hebbian Recurrent Network (HRN) is introduced.

Chapter 3

The Mechanism of Memory

Within the recent literature there has been considerable debate about the ability of backpropagation models to capture human memory phenomena. While there is a feeling that backpropagation models have much to offer, the results thus far have been mixed (Ratcliff, 1990; Lewandowsky, 1991). Backpropagation models provide mechanisms by which encodings, decision criteria and control functions can be learned as a consequence of exposure to the environment, yet, on some very basic variables like the degree of interference, they have not performed as well as conventional memory models.

Estes (1991) makes a telling comment in concluding his discussion of the difficulties of connectionist models:

> "A stray thought that comes to mind is that the situation is a bit reminiscent of a segment of the history of learning theory. In the learning theory of the period 1930-1960, it was assumed that the learning processes are basically the same at least for all of the higher animals and that the learning in the individual organism starts from a *tabular rasa*, the counterpart of a network of homogeneous and mutually connected nodes. During the next decade, however, under the impact of ethology and the beginnings of modern neuroscience, the prevalent view shifted to one that recognised biological constraints on learning ... and it is now quite generally assumed that learning in any organism, human or subhuman, builds on a substrate of species-specific predispositions and products of previous learning. Implementing this more biologically founded orienta-

> tion in connectionist learning models is a tall order; but the effort will surely have to be made sooner or later, and the results may cast some of the current problems in a new light."

(Estes, 1991, p 24)

Estes' position is supported by a large number of memory phenomena which do not seem to be directly reducible to environmental contingencies. For instance, the depth at which items are processed has been shown to have an effect on retention (Craik & Lockhart, 1972). In general, items processed semantically will be retrieved with greater accuracy than items processed either orthographically or phonologically [1]. It would seem likely that the nature of the information processing device plays a decisive role in distinguishing these different forms of processing.

Similarly, the impact of imagery on retrieval would seem to be more acutely dependent on the nature of the mechanism of memory than the environmental statistics. Bower (1972) presented subjects with pairs of concrete nouns. One group were required to form images of the two objects interacting. The control group were left to their own devices. Performance was superior for the group who formed images. The difficulty that this phenomena presents for an environmental account is that the nature of the input (i.e. the lists of words) was identical for both groups. The only difference was in the instructions which presumably affected the nature of the processing undertaken by the mechanism of memory.

The purpose of this chapter is to *start* to address what forms of bias must be added to a network to make the learning tasks commonly solved by humans tractable and to ensure that the network performs these tasks in a psychologically plausible way.

The discussion begins by outlining some of the contributions that backpropagation architectures can make to the modelling of memory. Next, some of the major aspects of memory phenomena that remain obstacles for backpropagation models are examined. Finally, the Hebbian Recurrent Network (HRN), which integrates learning and memory models, is presented.

[1] See Morris, Bransford, and Franks (1977) for an exception.

3.1 Learning Issues

In an interactive (learning) model, it is the interplay of the environment and the architecture that leads to performance (Bickhard, 1991). For example, in backpropagation models, performance is determined both by the architecture (i.e. structure of the interconnections, the transfer function, and the values of the parameters) and the statistical contingencies embodied by the training set.

The advantages of such an interplay can be examined in terms of the components of the memory system, which were identified in the definition of learning provided in the introduction:

> **Learning:** A change in behaviour due to the formation of representations, decision criteria and control processes as a consequence of the interaction of a mechanism and the environment.

In the following subsections, the learning of representation (subsection 3.1.1), decision criteria (subsection 3.1.2) and control (subsection 3.1.3) are considered in turn.

3.1.1 Learning Representation

The formation of appropriate encodings for items that will be entered into memory has been a difficult and largely unexplored area in the memory literature. While it is known that items with similar meanings and perceptual forms interact with each other to a greater extent than unrelated items (Osgood, 1949; Martin, 1965), there has been little progress in determining how the formation of such an encoding landscape may come about. Eich's (1982) presentation of the representations used in CHARM is indicative of the assumptions that are commonly made in the memory literature, at least among those who posit distributed memory systems. Items are represented by vectors of "abstract features", and the similarity of representations is assumed to be a consequence of overlap of these features. Typically, these representations are strongly constrained by the requirements of the mechanism in which they will be used. Eich (1982) assumes statistical independence of the representations of unrelated words, with the dot product being the measure of similarity of items. The dot product is a particularly good choice in the context of

the convolution/correlation mechanism of CHARM as it means that stimulus and response generalisation are immediate consequences. However, since the nature of the component features is deliberately left unspecified, there is no way to independently decide upon the values of the inter-item dot products, and typically these are free parameters, which are optimised to fit the empirical data. The arbitrary nature of the assumptions that underpin current mathematical memory models, including those concerning internal representation, has been criticised (Watkins, 1990).

Backpropagation models are able to construct internal representations (Rumelhart, Hinton, & Williams, 1986), thus offering a way of avoiding many representational assumptions. Typically, the hidden unit representations are formed as a consequence of both input similarity (e.g. the words "been" and "bean" might assume similar representations since they have similar orthographics and identical phonology) and functional similarity (e.g. the words "idea" and "concept" might adopt similar representations because they are used in functionally similar ways). Elman (1989, 1990) has demonstrated how a network in which the input representations have no similarity structure can exploit the functional similarity in the statistics of the training set to create a similarity landscape on the hidden units. Hence, backpropagation networks introduce a principled way in which "abstract features" might be formed as a consequence of the environment with which they are faced. The first of the design criteria for a model of memory is that internal representation be learned.

3.1.2 Learning Decision Criteria

Another aspect of the memory task that is usually held constant is the decision mechanism. For example, consider recognition as modelled by SAM, Minerva II, TODAM, CHARM and the matrix model. A signal detection framework is applied. Each model calculates a global matching strength (Humphreys et al., 1989) to be compared against a criterion (see section 1.3). In general, the criterion is assumed to be flexible allowing for some adaption, but the form of the matching strength function remains the same. There is no sense in which the matching function could be said to have been acquired.

In the majority of memory paradigms, however, subjects become more accurate

as they gain experience. While some of the improvement may be due to the refinement of the representation, a portion is attributable to an improvement in the ability to decide upon a response (Postman, 1969). Hence, the second design criteria of the HRN is that it be able to acquire decision criteria.

3.1.3 Learning Control

While the representations used in memory and the decision criterion that determine performance are well developed topics within the memory literature, the nature of the control processes is often left unelaborated (see Atkinson & Shiffrin, 1968, for an exception). For instance, how is it that the subject decides to give a response only during the test phase? Current mathematical models of memory assume that the answer to this question is embedded in the program or is part of the metamemory. Models that attempt to account for learning allow parameters on already existing processes to change rather than have these processes acquired through experience (c.f. Murdock, 1987). In order to approach questions about the nature of rehearsal, elaborative processing, imaging and other metamemory skills it would be advantageous to have a more exacting theory about how simpler control processes, such as when to output a response, are achieved. Consequently, the third of the design criteria is that control processes be acquired.

3.2 Memory Issues

In the previous section, the aspects of the memory system that might be acquired by a learning system such as a backpropagation network were outlined. To be serious alternatives to current models of memory, however, there are a number of criteria on which current memory models perform well that must be fulfilled. These memory criteria include maintaining significant capacity without introducing unrealistic amounts of interference (subsection 3.2.1), generalising to unseen lists of items using small numbers of training examples (subsection 3.2.2) and the ability to establish memory traces rapidly (subsection 3.2.3).

3.2.1 Capacity and Interference

The problem of catastrophic interference has received a great deal of attention in the recent literature (Ratcliff, 1990; McCloskey & Cohen, 1989; Lewandowsky, 1991; Brousse & Smolensky, 1989; Hetherington & Seidenberg, 1989; Wiles & Phillips, 1991; Kruschke, 1992; Chappell & Humphreys, in press, 1993). The difficulty arises when what has been learned is disrupted dramatically by subsequent learning. That is, there is too much retroactive interference. The problem is of particular importance in the modelling of recognition memory where the capacity is very large and the degree of interference is small. Certainly, it is not nearly as marked as in standard feedforward backpropagation architectures (Ratcliff, 1990).

Within the literature two major strategies have emerged in order to deal with the problem of catastrophic interference. The first involves increasing the orthogonality of items that are to be learned in succession. Lewandowsky (1991), Kruschke (1992) and French (1991) have suggested methods for encouraging orthogonality and, hence, decreasing the amount of interference within feedforward networks.

The alternative approach has been to use recurrent architectures to encode lists of items rather than single items on their hidden units (Nolfi, Parisi, Vallar, & Burani, 1990; Wiles & Phillips, 1991)[2]. The network has the task of learning a single higher order encoding function rather than a sequence of items and, hence, the interference is reduced. Unfortunately, this encoding function becomes much more difficult to acquire as the number of items to be encoded increases. Consequently, current recurrent architectures have severe capacity restrictions. The first of the memory criteria, then, is that the model avoid catastrophic interference while maintaining capacity.

3.2.2 Generalisation

Another issue closely related to interference is the degree of generalisation. With what proportion of the entire space of possible inputs must the network be presented to perform well on unseen cases? In the context of memory the important variable

[2]Brousse and Smolensky (1989) used a somewhat similar approach, in that lists of items were encoded in patterns of activation rather than in the weights. However, their approach employed a tensor product mechanism to construct the patterns, rather than relying on a recurrent architecture to learn an appropriate encoding scheme.

is the size of the vocabulary. Subjects have extensive vocabularies, yet are able to perform memory tasks involving any of the items within that vocabulary despite having limited experience with lists of these items. The second of the memory criteria, then, is that the model be capable of generalising on the basis of a very small proportion of the input set.

For traditional memory models, such generalisation is not a problem since the memory mechanisms are chosen so that they will operate accurately independently of which items are to be encoded. In the recurrent network architectures that have been applied to memory phenomena (Nolfi et al., 1990; Wiles & Phillips, 1991), however, there is no such constraint. The network must learn to recognise each new item. In addition, in the serial recall task investigated by Wiles and Phillips (1991), a significant proportion of the possible orderings of the items must also have been encountered. Brousse and Smolensky (1989) highlight this issue and suggest that the process of building the list representation be hard wired by using a tensor product of the items position and the item vector, effectively concatenating the list items. In order to fulfill the second memory criterion some such approach must be adopted.

3.2.3 Rapid Binding

The last of the memory criteria revolves around the distinction between memory and learning (Melton, 1963). Models such as that of Ratcliff (1990) assume that memorisation is best modelled by the learning processes of feedforward networks. Memory, however, seems to involve the rapid binding of already established representations rather than the acquisition of new representations (Wiles & Phillips, 1991). Enduring memories can be laid with presentation durations of just a few hundred milliseconds - not much time for a learning mechanism to operate. The time course of learning-to-learn effects, in contrast, tends to be in the order of minutes or hours. Furthermore, the development of metamemory skills (e.g. realizing that recall is more difficult than recognition) happens over several years. A model of memory should be capable of explaining the difference between memorisation and learning, and be able to account for the difference in the time scales.

3.3 The Hebbian Recurrent Network

3.3.1 The Objectives

In the last section, some of the important design criteria for a model that integrates learning and memory were discussed. In summary, an adequate model of human memory should be able to:

(1) Form representations that are sensitive to environmental statistics.

(2) Acquire decision criteria.

(3) Acquire control mechanisms.

(4) Avoid catastrophic interference and capacity restrictions.

(5) Generalise to lists of arbitrary order with large vocabularies.

(6) Form bindings on established representations rapidly.

Current backpropagation models can address the interactive issues but fail on the memory criteria on which traditional memory models do very well. By embedding a traditional memory model within a backpropagation learning model, all of these issues can be addressed. In the next sections, a learning model and a memory model are selected and then integrated to form the *Hebbian Recurrent Network (HRN)*.

3.3.2 The Model

Choosing a Learning Model

Memory phenomena involve temporal tasks. Furthermore, the information that is added to memory is affected by the current contents of memory. Memory is a closed-loop system in Murdock and Lamon's (1988) terminology (see also Lewandowsky & Murdock, 1989). While standard feedforward architectures are capable of forming representations and acquiring decision criteria, they cannot embody the temporal relations that characterise control problems.

The variety of temporal networks that have been developed fall into two classes[3]. The first class includes those networks that buffer input in order to maintain temporal information (e.g. *Time Delay Neural Network (TDNN)*, Waibel, Hanazawa,

[3]For an analysis of current temporal learning algorithms and their suitability in cognitive tasks see Regier (1992). In addition, Mozer (1993) provides a detailed classification of temporal network architectures.

Hinton, Shikano, & Lang, 1987; Waibel, 1989). In a TDNN, the input sequence is presented to the backpropagation network through a series of delays. Hence, if $i(t)$ is the input at timestep t, the network receives, $i(t), i(t-1), i(t-2), i(t-3), i(t-4)$ and $i(t-5)$ at the same time. In a memory task, the entire study list would be required by the network at the time of decision. While the depth of the network remains constant, the number of inputs grows both with the vocabulary and the length of the sequence. Not only is such an architecture prohibitively large, but it requires that each item be presented at each possible timestep within the training set. Only in this way can the TDNN learn to respond appropriately independently of the position of the item. Hence, a large training set would be required and the generalisation criterion would be violated.

The second class subsumes recurrent networks such as the Jordan network (Jordan, in press), the *Simple Recurrent Network* (*SRN*, Elman, 1989, 1990), *Back Propagation Through Time* (*BPTT*, Rumelhart et al., 1986) and *Real Time Recurrent Learning* (*RTRL*, Williams & Zipser, 1989). In contrast to the TDNN, which assumes that the input is buffered to maintain information, recurrent networks postulate recurrent connections through which information is cycled and, hence, preserved.

The recurrent connections can emanate either from the output units (e.g. Jordan network) or from the hidden units (e.g. SRN, BPTT, RTRL). There are, however, problems that are not able to be solved by the Jordan network because the information required to be maintained is not present at the output and is necessarily lost (Cottrell & Tsung, 1993; Dell, Juliano, & Govindjee, 1993). In particular, memory control paradigms such as rehearsal can operate without the requisite information in the outputs. Hence, to avoid restrictions on the control paradigms that can be implemented, hidden unit recurrency is required.

Having established the form of the architecture, it remains to choose the learning algorithm which will be applied. The BPTT algorithm unfolds a network (Rumelhart et al., 1986) creating a level for each timestep with tied weights between the timesteps. Because very deep networks are constructed by this method training is often very difficult and time consuming. In addition, a great deal of memory is required. The RTRL algorithm follows the same gradient as a BPTT network that

is unfolded for the entire length of a sequence, but does not require time or space proportional to sequence length. It is an $O(n^4)$ algorithm[4] (where n is the number of nodes) and is slow for large architectures.

The Simple Recurrent Network (SRN, Elman, 1989, 1990) can be thought of as an approximation to the fully recurrent networks such as BPTT and RTRL. Instead of backpropagating over the entire length of the sequence, it considers only the last timestep. While being faster (i.e. $O(n^2)$) than either BPTT[5] or RTRL, it does not actively seek to maintain relevant information, but makes use of the information that is maintained by chance. Because the current state is a consequence of the preceding states, it is possible that relevant information will be retained until needed. In practice, this information is often lost before such time as the error signal can be generated. Without useful information in the form of hints being required on the outputs, the ability of the SRN to remember is limited to short study lists (Phillips, 1991).

Despite its limited nature the SRN performed well on the language prediction tasks to which it was applied (Elman, 1989, 1990). The hidden unit patterns that were chosen by the network formed a similarity landscape that was related to syntactic and semantic structure in the training corpus. In addition, the SRN developed the ability to decide upon a response (or response set) given the prior context. Furthermore, the network was able to retain information that could influence later encoding and predictions, such as pluralisation, over short time spans - a limited form of control. Because the SRN fulfills the learning criteria established earlier it was chosen as the learning mechanism of the HRN.

Current recurrent network models of human memory (Nolfi et al., 1990; Wiles & Phillips, 1991) attempt to learn to build a representation of the entire list. This representation requires training on a significant proportion of all possible lists in order to ensure generalisation (Wiles & Phillips, 1991). People, however, do not

[4]At least in its original formulation (Williams & Zipser, 1990), although an $O(n^3)$ algorithm has recently been developed (Schmidhuber, 1992).

[5]This statement needs some qualification. There are many variations on BPTT. If the SRN is compared against BPTT where error is backpropagated at the end of each sequence rather than after each pattern, the time complexities per sequence are equal. However, the number of weight updates is decreased from l, where l is the length of the sequence, to just one per sequence. Experience with feedforward nets suggests that if there is redundancy in the training set online training is often faster than batch because of the increase in the number of weight updates (Hinton, 1993).

receive such extensive training. To perform the sorts of tasks people do routinely, a model must include a form of bias that circumvents this learning problem, that is, the model must learn to encode items not entire lists. While the recurrent activations are the only means by which information can be stored, however, they will continue to be used to store the sequence of items. One solution is to provide a separate mechanism that is capable of storing lists. Fortunately, such a mechanism has been the subject of research into human memory for several decades. In the following section, this literature is examined to find a mechanism suitable for inclusion into the SRN.

Choosing a Memory Model

From the extensive work in the verbal learning tradition during the 1960s and 1970s emerged a series of mathematical models that were influential during the 1980s (Raaijmakers & Shiffrin, 1981; Murdock, 1982; Eich, 1982; Pike, 1984; Hintzman, 1984). These models have made significant contributions to the precision with which memory models are now specified and have helped to focus research (Hintzman, 1991).

With the exception of Search of Associative Memory (SAM Raaijmakers & Shiffrin, 1981) these models employ distributed representations and are all possible candidates for incorporation into the SRN. Minerva II (Hintzman, 1984) is an exemplar model, meaning that each memory item is stored separately. Such a model is able to account for situational (Hintzman & Block, 1971; Greene, 1992) and categorical (Greene, 1992) frequency judgements, but incurs the cost of storage proportional to time. The alternative is to employ a blend memory in which images are superimposed upon each other. TODAM (Murdock, 1982), CHARM (Eich, 1982) and the Matrix model (Pike, 1984; Humphreys et al., 1989) are examples of superposition models, and were considered to be more feasible for implementation in the SRN. The final choice was between the correlation/convolution methods of TODAM and CHARM, and the Hebbian scheme of the Matrix model. While the Hebbian rule has well established roots in the connectionist literature (Anderson, Silverstein, Ritz, & Jones, 1977; Hopfield, 1982), the correlation/convolution method has also been studied in this context (Plate, 1991, 1992). Due to the arguments put forth by

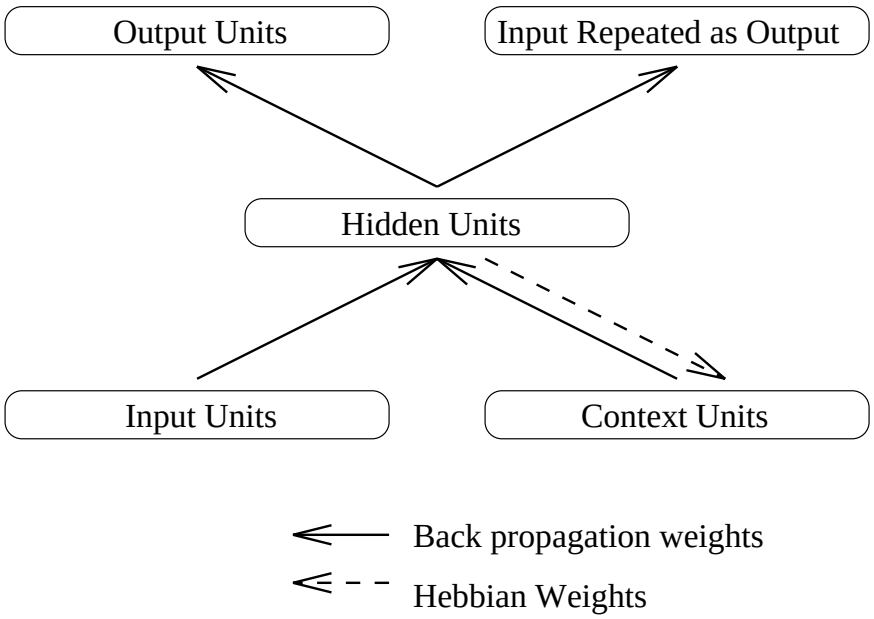

Figure 3.1: The Hebbian Recurrent Network (HRN). The solid arrows are sets of weights that are modified using the backpropagation algorithm. The dashed line represents the feeding of hidden unit activations through a set of Hebbian weights to the context units in preparation for the next timestep. The Hebbian weights are updated after the activations are fed through. In addition, the input pattern was required on the output to ensure that the states corresponding to different inputs were separated.

Pike (1984) that the matrix model is less subject to noise, and the obvious mapping of the cells of a matrix to the weights of a Hebbian network, the Matrix model was chosen. Allowing weights to store information instead of requiring the activation pattern to encode it was expected to improved performance in a plausible fashion (Schmidhuber, 1991; Mozer, 1992).

Description of the Hebbian Recurrent Network

There are two ways in which one may view the HRN: The first is to regard it as a matrix memory to which optimised weights have been added to allow the acquisition of the control, decision and representation aspects; the second way is to think of it as an SRN with longer term memory.

The architecture of the HRN is similar to that of the SRN, in that the input and context layers are completely connected to the hidden layer and the hidden layer is completely connected to the output layer (see figure 3.1). In contrast to the SRN, which copies the contents of the previous hidden layer to the context layer, the hidden layer of the HRN is connected to the context layer by a set of Hebbian

The HRN Algorithm
1. Set inputs. Context activations and Hebbian weights are set to zero before each sequence.
2. Feed activation through the input and context weights to the hidden layer and from the hidden layer to the outputs using the symmetric sigmoid (Stornetta & Huberman, 1987). $$f(X) = \frac{2}{1+e^{-X}} - 1$$ $$A_H(t) = f(A_I W_{IH} + A_C W_{CH})$$ $$A_O(t) = f(A_H W_{HO})$$ where A_I are the input activations, A_H are the hidden activations, A_O are the output activations, W_{IH} are the input to hidden weights W_{CH} are the context to hidden weights and W_{HO} are the hidden to output weights.
3. Use the usual backpropagation rule to train the input-to-hidden (W_{IH}), context-to-hidden (W_{HC}) and hidden-to-output (W_{HO}) weights.
4. Calculate the values for the context units to be input at the next time step by multiplying the hidden activation pattern by the Hebbian matrix and dividing by a Scaling Factor (γ). $$A_C(t+1) = f(A_H(t) W_{HC}(t)/\gamma)$$ The scaling factor was set to be 0.05 * the number of hidden units
5. Update the Hebbian network by autoassociating the hidden pattern. $$W_{HC}(t+1) = (1-\lambda) W_{HC}(t) + \lambda A_H(t) A_H^T(t)$$ where W_{HC} are the hidden to context weights and λ is the memory decay.

Table 3.1: The HRN algorithm. Note that the scaling factor value was introduced to stop the context units from saturating. When updating the Hebbian weights, retaining $1 - \lambda$ and adding λ of the new vector ensures that weights are bounded (since the activations are bounded between 1 and -1), and that the mechanism will be stable. The Hebbian weights are updated during both training and testing.

weights. It is these weights that form the memory of the system. At any given timestep, the context layer contains the results of the last memory retrieval. In all of the experiments reported in this thesis, additional output units were added that were required to reproduce the inputs.

The HRN algorithm is also similar to that of the SRN (see table 3.1). Before each sequence the context unit activations and the Hebbian weights are set to zero. At each timestep the hidden units receive input from both the input units and the context units. The outputs of the network are calculated on the basis of the hidden patterns. The activation function is the symmetric sigmoid. In addition, the hidden patterns are used to probe the matrix memory. The result of this probe is instantiated on the context units in preparation for the next timestep.

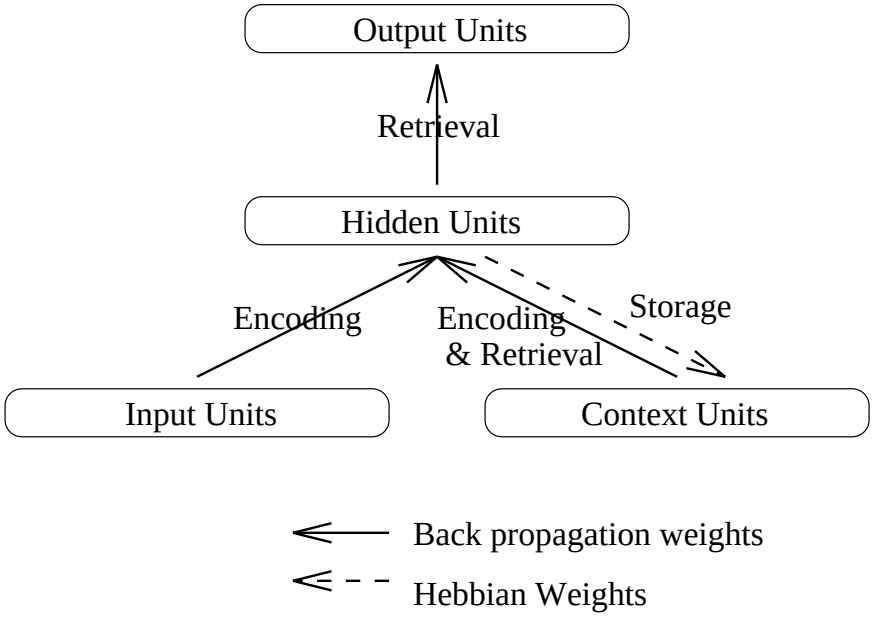

Figure 3.2: Memory processes in the HRN. The input to hidden and context to hidden weights implement the encoding process. The hidden to output and context to hidden weights implement the retrieval process, while the Hebbian weights from hidden to context store the memory traces.

The weights are updated after the hidden and output activation patterns have been updated, but before the context units are changed. The backpropagation weights are updated using the backpropagation rule as in the SRN. The Hebbian weights are updated by autoassociating the hidden unit activation pattern. The outer product of the hidden vector with itself is calculated and added to the existing Hebbian matrix.

Figure 3.2 outlines where the processes involved in memory tasks are implemented in the HRN. The input-to-hidden and context-to-hidden weights implement the *encoding process*. The context-to-hidden and hidden-to-output implement the *retrieval process* and the Hebbian weights from the hidden-to-context units are responsible for storage.

Rationale of the HRN

The objective in designing the HRN was to satisfy the learning and memory criteria discussed in section 3.2 and listed in subsection 3.3.1. The Hebbian weights have been added to accommodate the memory criteria and, provided the representations to be retained are available to matrix memory, it should be possible to satisfy these criteria. The learning criteria, however, impose two important constraints on the

structure of the HRN.

Firstly, the representations should be determined by the dynamics of the network as a consequence of learning and not chosen *a priori* by the experimenter. Hence, the inputs to the memory system come from the hidden activations of the backpropagation network. The representational scheme will be formed as a result of the input and required outputs, that is, as a function of the environment in which the network must operate. Furthermore, it is also necessary that the the matrix memory be autoassociative. The problem with using a feedforward Hebbian memory in a system in which the representations are learned is that the Hebbian memory itself is solely responsible for generating the representation at its outputs. Any Hebbian learning could only reinforce the output patterns that already occurred as a consequence of the initial weights. For instance, if the initial weights were all zero the only outputs that would be generated would also be zero. Using a Hebbian weight update scheme, however, the zero outputs would result in no change to the weights and hence nothing could be retained. Hence, the Hebbian system is necessarily autoassociative, it necessarily has a hidden layer as input, and that hidden layer must feed to and from error correcting weights.

The splitting of the Hebbian and backpropagation weights is not intended to suggest that these must represent different sorts of weights in the brain. What is being identified is the functional disparity. The Hebbian units are responsible for storage while the backpropagation weights perform function approximation. It may be the case, however, that both of these processes can be captured by a single weight update rule[6].

The second point is that the outputs of the Hebbian memory are available at the input to the feedforward backpropagation architecture. Consequently, the result of probing memory can be used to construct the next cue to memory, allowing the control aspects of the task to be acquired. In addition, such a recurrent system makes chains of recollection possible. The ability to perform these chains of recollection is a prominent aspect of everyday recall (Lewandowsky & Murdock, 1989).

At this point, the role of the "context" units may need some clarification. This terminology has been borrowed from Elman (1989, 1990) and does not correspond to

[6]Steven Phillips is currently attempting to integrate Hebbian and error correcting notions into a single update rule.

the context cues that are often featured in models of human memory (c.f. Gillund & Shiffrin, 1984; Humphreys et al., 1989; Chappell & Humphreys, in press, 1993; Heathcote, 1993). The context units receive the outputs of memory and, hence, form the "context" of the next input to the memory system. The context cues used in the memory literature are derived from the experimental instructions and define the episode from which a subject is expected to recognise items. In the following simulations, change of context has been implemented by resetting both the context units and the Hebbian weights. This mechanism decreases simulation time, but is insufficient for tasks that involve retrieval from multiple lists (i.e. multiple contexts). To model these tasks, context could be included as part of the input. As is often done in distributed memory systems, each episode could be represented by a vector and the appropriate vector would be reinstated at the time of test to indicate from which context the network is required to retrieve.

3.3.3 The Task: Episodic Recognition

To test the performance of the HRN the episodic recognition paradigm has been chosen. In this paradigm, a list of words is presented to the subjects during a study phase. Subsequently, a series of words are shown, some of which occurred in the study list and some which did not. Subjects are asked whether they have seen each item within the study context. Human subjects are surprisingly accurate on such a task. For instance, Shepard (1967) gave subjects a study list of 540 words and found that they were able to recognise 88% of the words. Hence, one of the questions that will be addressed in the following simulations is how the performance of the HRN declines as the list length increases (see section 3.2.1). Furthermore, each of the words used in Shepard's study was taken from a population of 600 English nouns and adjectives - a substantial vocabulary. A second issue of interest that derives from the experimental results is how the performance of the HRN decreases as the vocabulary size increases (see section 3.2.2).

3.3.4 Stepping through a Simple Example

Table 3.2 outlines how the task was mapped onto the network. On the input, each of the study items was presented one at a time. These were followed by a single

	First Study Item	Second Study Item	Third Study Item	Probe	Answer
Input	A	B	C	B	Blank
Targets	Blank	Blank	Blank	Blank	Yes

Table 3.2: An example input/output sequence for the episodic recognition task.

probe item and then by a "Blank" signal. On the output, the network was required to respond with the "Blank" symbol on all occasions except during the answer phase at which time the network was required to respond with either "Yes" or "No" depending on whether the probe pattern occurred in the study list. Local encodings were used throughout and the non zero entries were set to 0.7 (to avoid the low gradient tails of the sigmoid).

Figure 3.3 shows the HRN applied to a small recognition task. In this case, the study lists consist of two items. At the third timestep an item is input to be recognised and at the fourth timestep the recognition decision is made. The diagrams on the left hand side demonstrate the processing of a target item (i.e. an item that was present in the two item study list) and the diagrams on the right hand side show the processing of a distractor item (i.e. one that was not present in the two item list). There are four input patterns, one for each of the three items as well as one "Blank" symbol, which is input when the recognition decision is to be made. At the output there are three patterns. For the first three timesteps the "Blank" pattern is output demonstrating that the network has successfully learned not to "babble" when no output is expected. The other outputs represent "Yes" and "No", the possible responses to the recognition decision. Four hidden units were used and, consequently, there were four context units. In addition, the inputs were repeated at the output (giving four more outputs). These extra targets were added to condition the error space during learning and do not affect post training processing.

The backpropagation weights of the network were trained to perform recognition by presenting study/test sequences. Once training was complete the backpropagation weights (but not the Hebbian weights) were frozen.

Consider the processing of the target, as on the left hand side of figure 3.3. The items of the list (i.e. "A" and "B") are presented to the inputs, and with the context

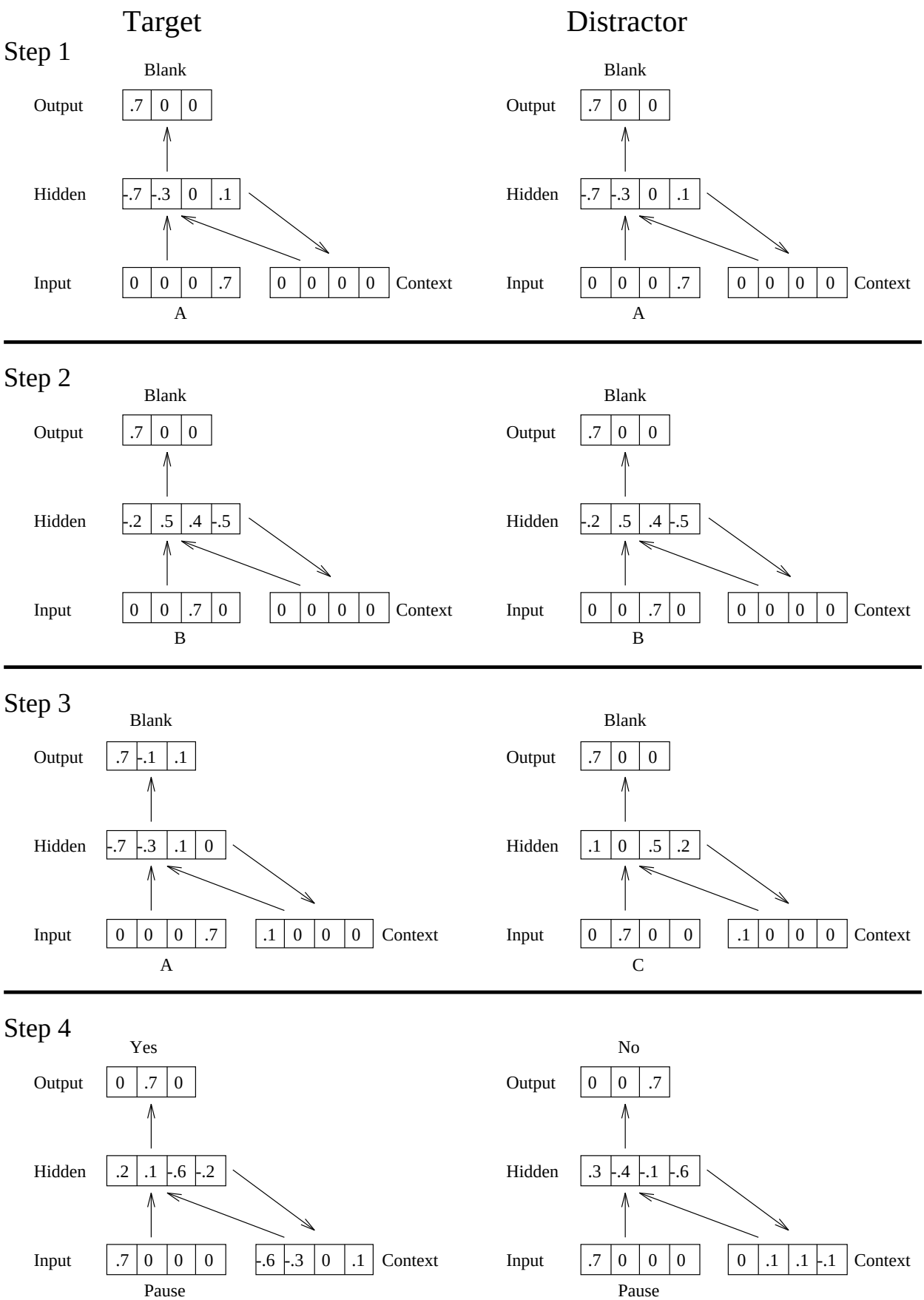

Figure 3.3: The trained HRN processing a target and a distractor. See section 1.3.4 for an explanation.

(which is zero to begin with) are used to form a cue. Because the Hebbian weights (which will be refer to as *medium term store*, MTS) are zero, there is no activation of the context units. If the patterns at the hidden units in response to these two inputs are denoted by A_H and B_H respectively, then, at the end of processing of the study list the Hebbian matrix is $(1-\lambda)\lambda A_H A_H^T + \lambda B_H B_H^T$ where λ is the memory decay parameter.

At step three, the probe item is input. Apart from some noise, the context units are zero and, hence, the hidden units will contain A_H. In the example, the hidden pattern is (-.7,-.3,0,.1). When this pattern is multiplied by the Hebbian weights there is a match:

$$A_H^T((1-\lambda)\lambda A_H A_H^T + \lambda B_H B_H^T) = (1-\lambda)\lambda |A_H|^2 A_H \tag{3.1}$$

since $A_H B_H^T \approx 0$. In this example, the context units have the pattern (-.6,-.3,0,.1). At step four, the network determines if the context units represent a valid item pattern. In this case, they do and the network responds with "Yes".

Now consider the distractor example on the right hand side of figure 3.3. The study items are input in the same fashion. When the "C" item is input as the probe item, however, there is no such pattern in the MTS and, hence, the context units are near zero (c.f. 0, .1, .1, -.1) when the recognition decision is to be made. The network has no trouble determining that this item has not been seen and responds "No".

While this architecture is capable of performing recognition tasks, it is not in principle restricted to such tasks. In general, the current input and the internal state are used to form a cue for MTS. Such a cue will recall any previously stored pattern that is either similar or the same as itself. Hence, presenting a concept is likely to recall other related concepts. Once these concepts have been recalled they participate in the formation of the next cue set for MTS and so on.

3.4 Simulation Method

The SRN and HRN were applied to the episodic recognition task[7]. The recognition task was mapped onto the network in the manner described above (see table 3.2). The effect of list length[8] and number of items in the training set was explored. In addition, the effect of the size of the vocabulary and the magnitude of the memory decay parameter on the performance of the HRN was investigated. All results are reported on 20 trials with a learning rate of 0.05 and 20 hidden units. Except in those simulations in which the memory decay is explicitly being varied, it was set to 0.2. Unless otherwise stated the vocabulary was set to twice the list length and the training sets consisted of 500 sequences (i.e. 500 * (listlength+2) patterns). The test sets consisted of 200 sequences. Each trial involved restarting the network with a new set of weights and new training and testing sets. The simulator was written in C on a SparcStation II platform and is available from the author on request.

Performance was measured at each decision timestep. The largest of the "blank", "yes" and "no" activation levels was taken to be the response of the network.

3.4.1 Measures

In the memory literature it is common to report performance in terms of *d'* (Swets & Green, 1961). The *d'* value is a measure of sensitivity derived from signal detection theory and is the distance between the noise and signal distributions divided by the standard deviation of the noise distribution (i.e. $\frac{\mu_{signal}-\mu_{noise}}{\sigma_{noise}}$). In general these values are not available, however, and it is commonly calculated using the number of times the subject correctly recognises an item from the study list (i.e. the hits) and the number of times they incorrectly recognise an item when it did not occur in the study list (i.e. the false alarms). A hit was scored when the network responded "yes" when the test item did occur in the study list. A false alarm was scored when either a "yes" or a "blank" response was given when the correct answer was "no".

[7]The RTRL algorithm was also attempted, but for the size of the training sets used the simulations proved too slow even on lists of length three. After almost two weeks running in the background on a SparcStation II only 300 epochs were complete. Thanks to Jeff Elman for the tlearn simulator on which the RTRL simulations were run.

[8]List length refers to the length of the study list, not the length of the entire sequence. Hence, if the list length is three there are five patterns in the sequence.

The value of *d'* was found by subtracting the z score for the false alarm rate from the z score for the hit rate (i.e. $z_{hit\ rate} - z_{false\ alarm\ rate}$).

Normally, when applying signal detection theory there are only two responses and a negative value of *d'* indicates bias against correct responding. In these simulations, however, a third response (i.e. "blank") is added and the chance baseline is reduced below 0.5 leading to negative *d'* values. However, the "blank" response was rarely given during the decision phase after the first few epochs of training. Hence, the problem quickly reduced to a two response paradigm and negative *d'* values were eliminated.

3.5 Results and Discussion

3.5.1 List Length

The performance of the SRN decreases very quickly as list length is increased and is at chance levels after only five items (see figure 3.4). In the HRN, the decrease in performance on lists of length three to eight is gradual. There is a small increase from lists of length eight to lists of length nine, but it is not significant, $F(1,38) = 0.563$, $p = 0.458$. This pattern is in general agreement with empirical findings, although no effort was made to match quantitative predictions. When the list length reaches 10, however, performance drops to chance. This aspect of the list length curve is clearly at odds with the empirical findings. However, in these simulations, the vocabulary size was set to be twice the list length so as to maintain a 0.5 probability of a positive test item. Hence, when list length reaches 10 the vocabulary has reached 20 items. Beyond 20 items it is impossible for a network that has 20 hidden units to maintain orthogonal hidden unit representations. The inability of the network to memorise larger lists is a consequence of the fact that the vocabulary size has reached the number of hidden units. Therefore, in a psychological model, it would be assumed that the number of hidden units would be much greater than both the size of the list and the size of the vocabulary, as is the case in TODAM, CHARM and the matrix model.

The learning curves for each of the architectures at each list length are shown in figures 3.5 and 3.6. In contrast to the SRN, performance rose very quickly in the

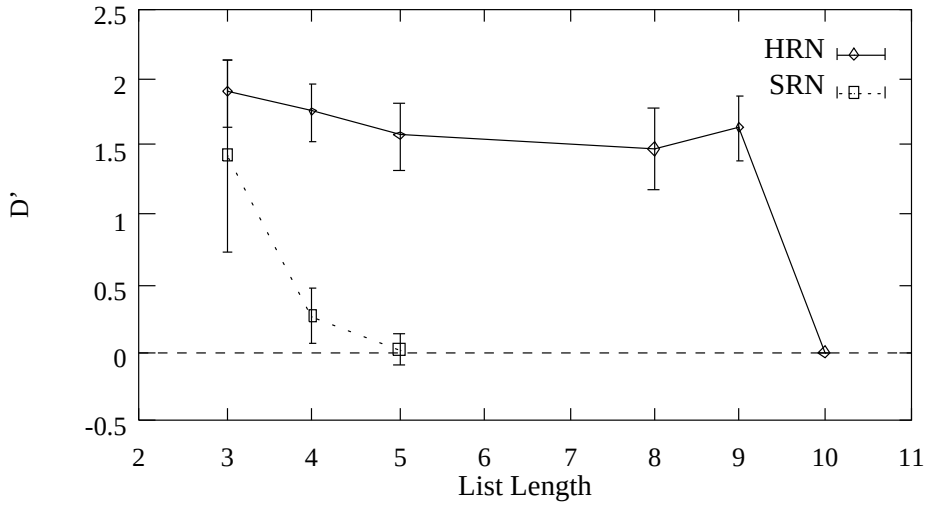

Figure 3.4: Performance (D') as a function of list length on the HRN and SRN. The HRN performed better than the SRN on lists of length nine or less. On lists of length 10 the HRN's performance falls to chance. The bars indicate the 95% confidence intervals.

HRN particularly at the lower list lengths. At the longer list lengths the increase in the number of items to be learned slows down the rate of improvement. In addition, the variances for the HRN graphs are much lower than those of the SRN for three item study lists, particularly in the latter stages of training. In the SRN there is no guarantee that the information required to make a recognition decision will be maintained. Because the true gradient through time is not being followed, a set of initial weights that does not maintain the requisite information sufficiently from the outset cannot be trained to do so. As training progresses, however, the performance on those simulations for which sufficient separation was maintained will continue to improve, leading to higher variances. In the HRN, the information required to make the recognition decision is maintained in the Hebbian weights. Furthermore, the retention mechanism is constant and the only thing that jeopardises the asymptotic performance is the orthogonality of the hidden unit patterns. For this reason, the variances in the HRN tend to be smaller.

Another interesting feature is the drop in performance that occurs in the HRN. For instance, on three, four and five item lists, performance reaches a maximum in the first 100 epochs and then decreases to an asymptote. The drop is not simply a consequence of overfitting. The training results show the same decrement in performance as the test results.

As suggested by Lewandowsky (1991) the symmetric sigmoid was used. Any

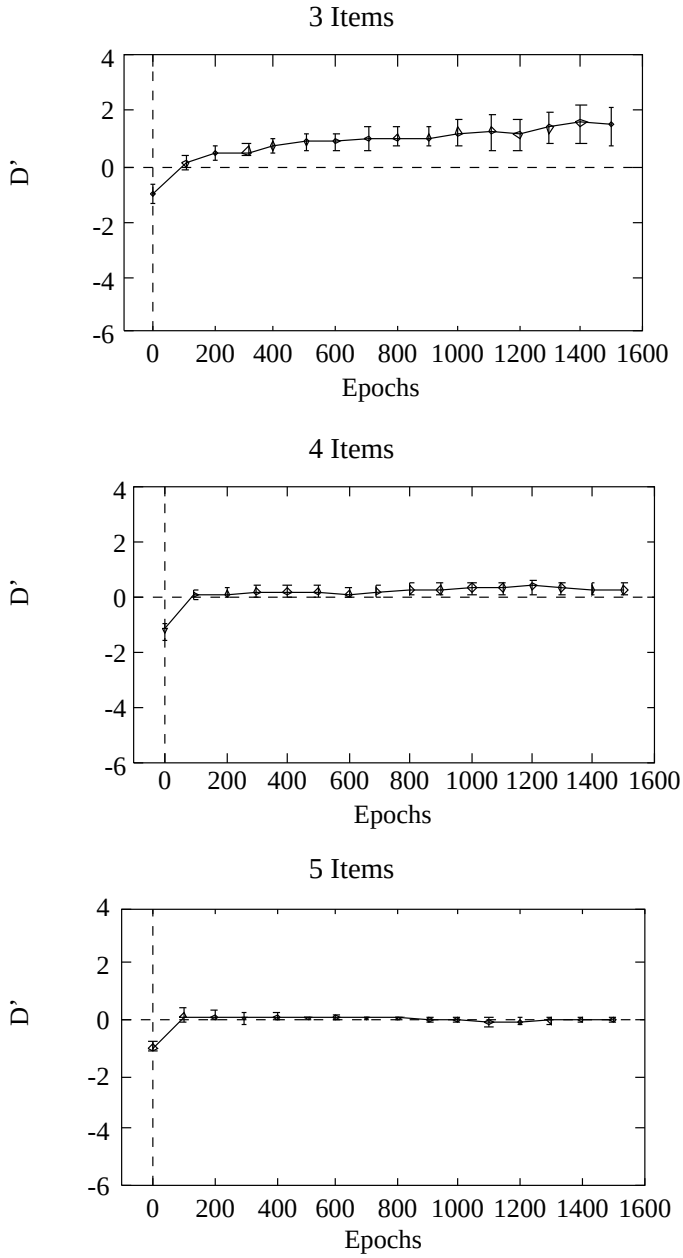

Figure 3.5: Learning curves for the SRN as list length was varied. The above graphs show the average *D'* (on the test set) as the number of epochs increased to 1500. The bars represent the 95% confidence intervals for the 20 trials. The only condition in which performance was consistently above chance was when three items were presented. The average *d'* value was significantly above chance by 200 epochs and increased gradually as training progressed. An interesting feature of the three item curve is the increase in the variances as training progresses. Some sets of starting weights will maintain the information necessary for the recognition decision and others will not. The SRN is not following the gradient through time but rather is reliant on the information that is maintained. Hence, if the information is not present from the start the SRN continues to perform at chance levels. On those occasions when the information is preserved, however, its performance improves. As training progresses, the difference between those simulations in which the information is retained and those in which it is not increases, and consequently, the variance increases.

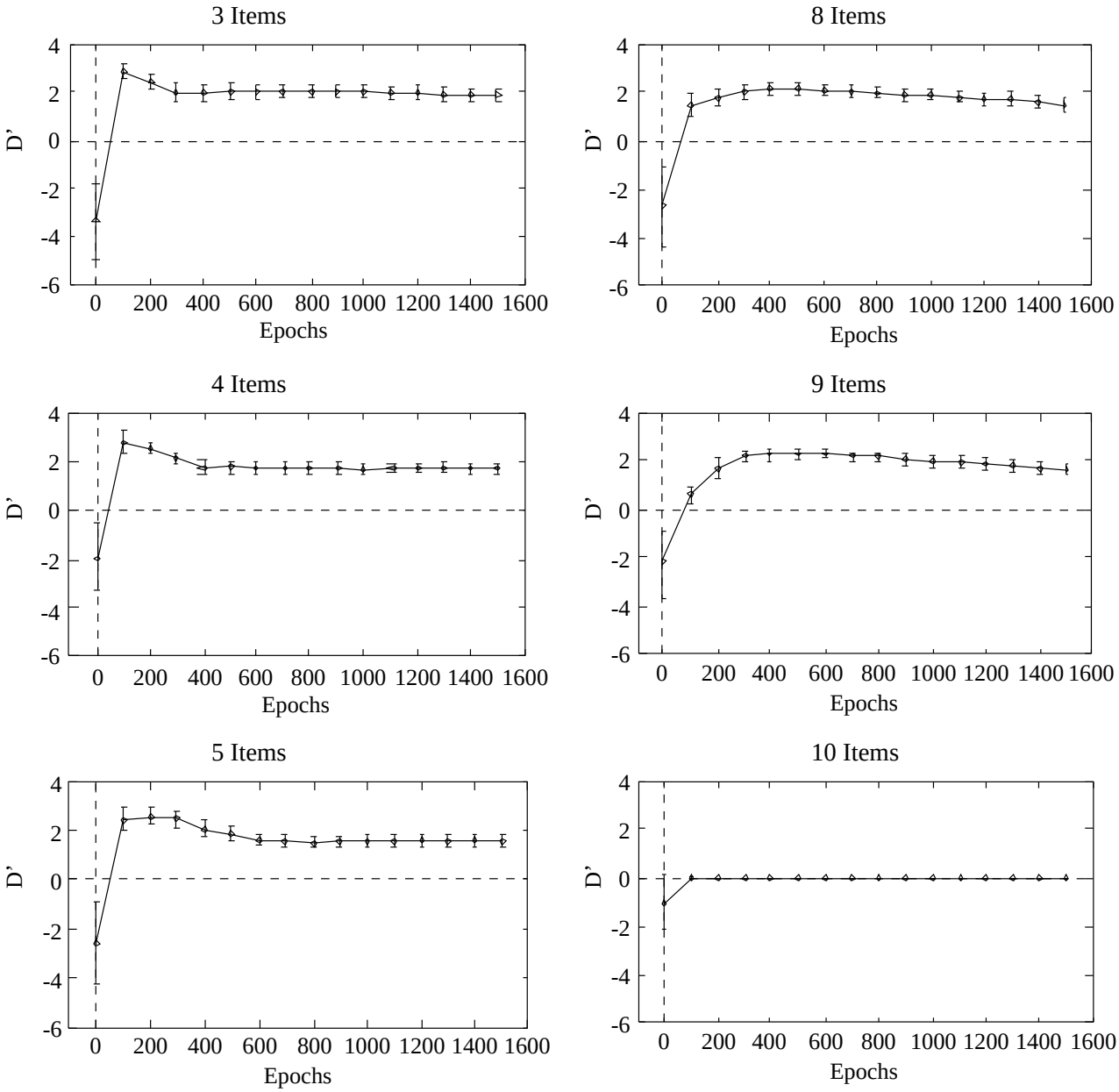

Figure 3.6: Learning curves for the HRN as list length was varied. The above graphs show the average D' (on the test set) as the number of epochs increased to 1500. The bars represent the 95% confidence intervals for the 20 trials. In contrast to the SRN, performance rose very quickly in the HRN, particularly at the lower list lengths. In all of the simulations except those involving the study lists of length 10, the average *d'* value was significantly above chance after 100 epochs. In addition, the variances for the HRN graphs were much lower than those of the SRN. Another interesting feature is the drop in performance that occurred in the HRN. In the three, four and five item lists, performance reached a maximum in the first 100 epochs and then decreases to an asymptote. It is important to note that the number of vocabulary items was maintained at twice the list length. Hence, the 8 and 9 item lists have 16 and 18 items to learn to autoassociate respectively which would account for their slower increases. The dramatic drop in performance for the 10 item lists occurs as a consequence of saturation of the hidden unit representational space. The items in a study list of length 10 were chosen from a vocabulary of 20 items. To use the Hebbian memory efficiently orthogonal hidden unit representations of the 20 items must be constructed. In these simulations, however, there were only 20 hidden units so that maintaining orthogonality is very difficult.

two vectors whose components are chosen from a uniform distribution in the range 1 to -1 will be statistically independent (i.e close to orthogonal). Lewandowsky (1991) demonstrates the decrease in catastrophic interference that ensues. However, unless large numbers of hidden units are used, mutual orthogonality of the pattern set quickly diminishes as the number of patterns increases. Correspondingly the performance of the Hebbian matrix drops.

In addition to the drop incurred by increasing the number of input patterns, the orthogonality of the hidden unit patterns decreases as training progresses (see figure 3.7). The measure of orthogonality used is:

$$Orthogonality = 1 - \frac{\sum_i^n \sum_{j<i} |\frac{v_i . v_j}{|v_i||v_j|}|}{n(n-1)/2} \tag{3.2}$$

where v_i is the average hidden pattern for input symbol i, and n is the number of input symbols. This measure is one minus the average normalised dot product. It ranges between one and zero and is at it maximum when all patterns are orthogonal to one another. The mutual orthogonality of the hidden patterns increases to a maximum after the first 100 epochs and then drops slowly to an asymptote as training progresses. It is this drop in the orthogonality that compromises the Hebbian memory and leads to the drop in performance of the network.

There are methods for encouraging orthogonality of hidden unit patterns (French, 1991; Chauvin, 1989) and it seems likely that these would improve performance.

3.5.2 The Memory Decay Parameter

The HRN was tested on the 4-item recognition problem with a range of memory decay values. Figure 3.8 shows the effect on performance as measured by *d'*. The performance increases to a maximum at a memory decay of 0.3 and then trails off as the decay rate is increased to 1. The optimum value is dependent on the length of the list. Figure 3.9 shows that as the memory decay parameter is decreased to zero the decay curve flattens. Hence, longer lists will perform optimally with smaller values of the memory decay parameter.

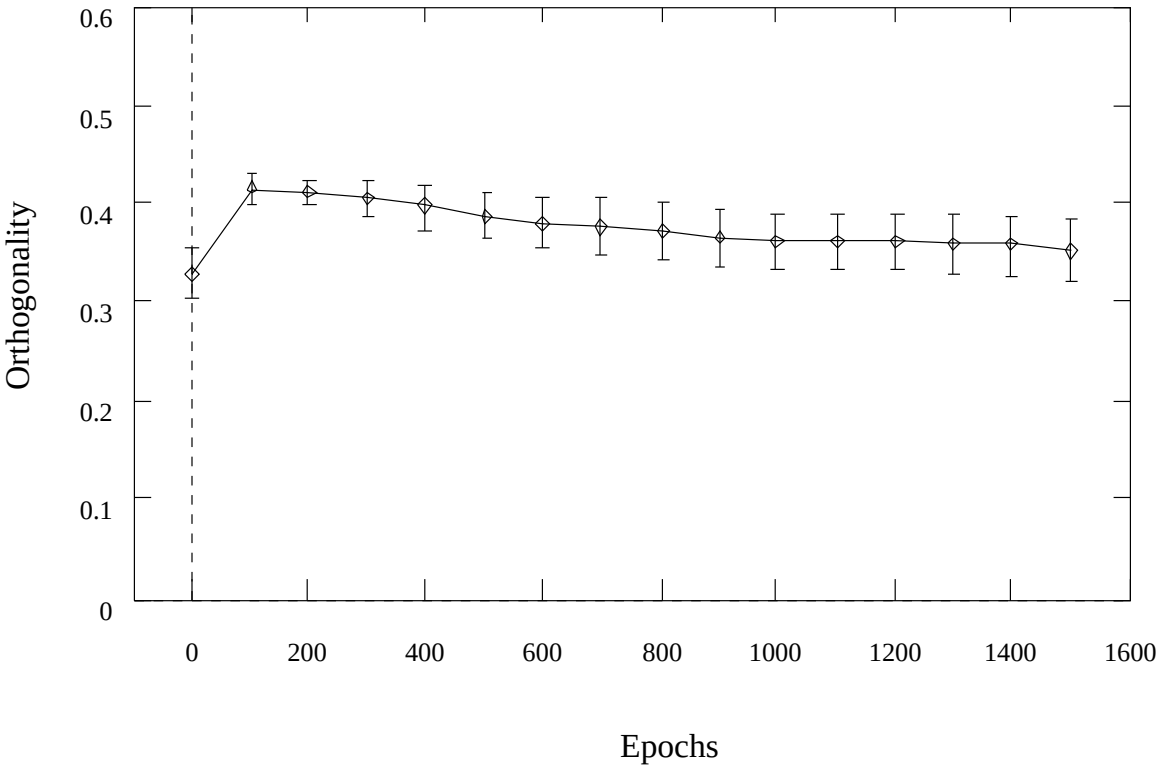

Figure 3.7: Orthogonality of hidden unit patterns of the HRN as a function of training time. See text for a description of the calculation of the orthogonality measure. The graph shows the average over 20 runs. The bars represent the 95% confidence intervals. As training progresses the mutual orthogonality of the hidden unit patterns decreases, jeopardising the Hebbian networks ability to perform accurately.

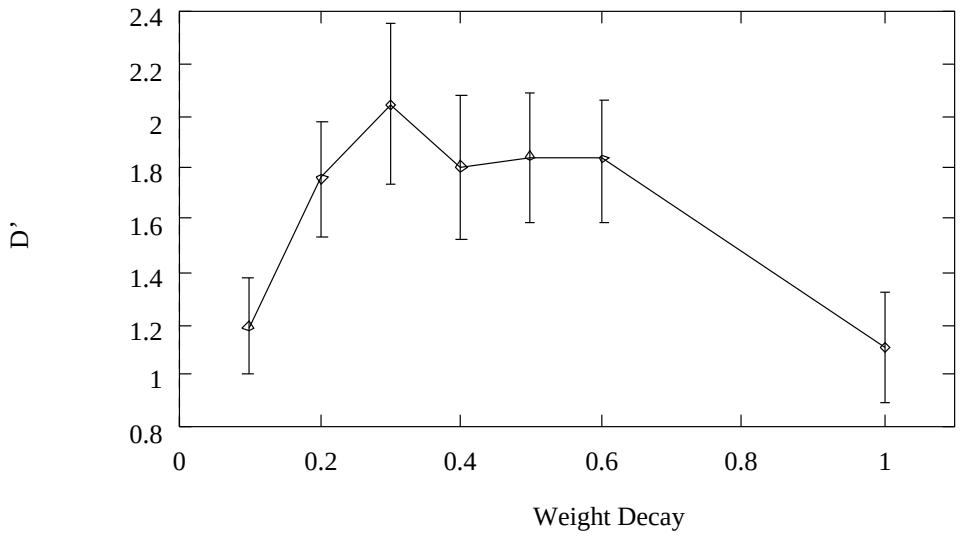

Figure 3.8: Performance (D') of the HRN as a function of the memory decay parameter on the 4-Item recognition task. Performance peeks at a memory decay of about 0.3 and drops as the decay increases to 1. The bars indicate the 95% confidence intervals.

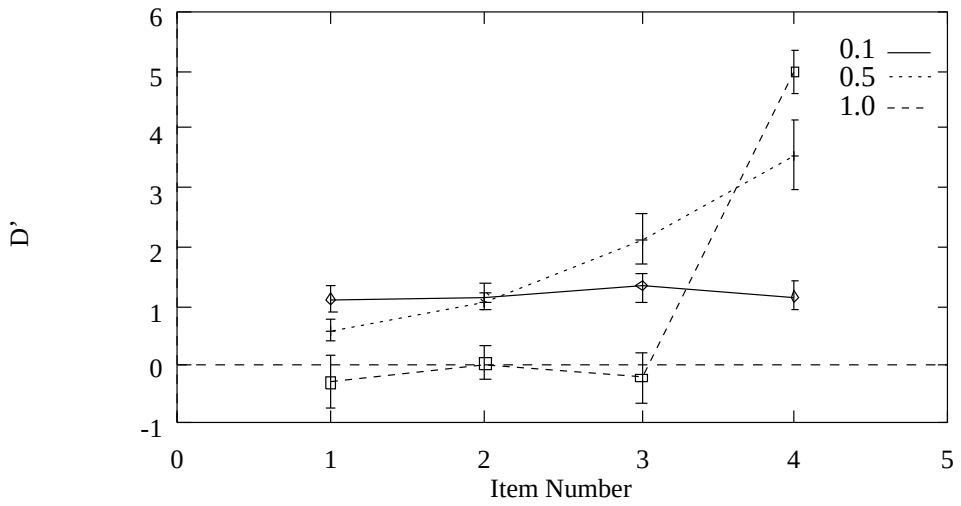

Figure 3.9: Performance (D') of the HRN as a function of item position for three values of the memory decay parameter. At a memory decay of 0.1 the new value of each weight is primarily a consequence of the previous value of that weight. Hence, position makes little difference and a flat position curve is observed. As the memory decay increases from 0.1 to 0.5 and from 0.5 to 1.0 the emphasis is moved towards the end of the list. In the extreme case of a memory decay of 1.0 there is no contribution from earlier timesteps and hence performance on the last item is very good while all other items fall to chance. The bars indicate the 95% confidence intervals.

3.5.3 Generalisation

One important criterion on which to judge the value of a network is its ability to generalise to unseen cases. In the current context, the HRN is required to respond correctly to test items when presented with novel study/test sequences[9]. As Geman, Bienenstock, and Doursat (1992) point out, there is a tradeoff between the generality of an architecture (degree of bias) and the number of training examples that will be required to train it. Backpropagation nets such as the SRN, BPTT and RTRL are approximations to unbiased systems, the approximation becoming more accurate as the number of hidden units is raised. Hence, they are prone to poor generalisation in the absence of significant numbers of training examples[10]. Figure 3.10 demonstrates the generalisation of the HRN as compared to the SRN. At 200 training examples the training performance of the SRN is much better than that of the HRN. The test performance, however, is not as good. The SRN is not generalising to unseen cases

[9]Note that in these simulations all of the items would have been presented during training. What would not have been seen are the sequences of items in the test set.

[10]In the feedforward case, feasible generalisation on combinatorial domains can be achieved by bounding the number of hidden units (Phillips & Wiles, 1993), however, the recurrent case remains an open question.

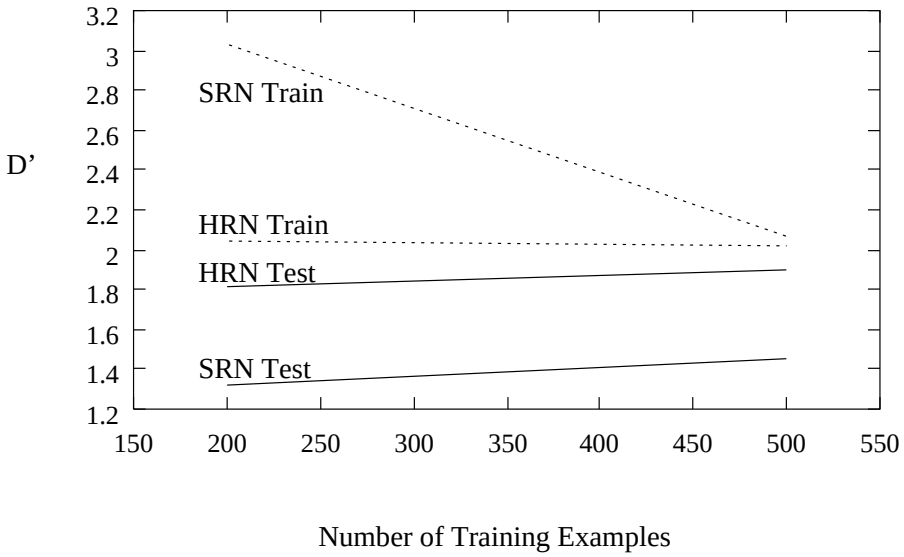

Figure 3.10: Performance of the SRN and HRN as a function of number of training examples. The HRN incorporates more bias than the SRN. Hence, as the number of training sequences decreases from 500 to 200 the change in the disparity between the training and testing performance is much greater for the SRN than for the HRN.

as well.

3.5.4 Vocabulary Size

Adding vocabulary items to the SRN tends to make learning much more difficult. Human subjects on the other hand seem to maintain very large vocabularies with surprisingly little interference or unlearning. Figure 3.11 shows performance as the vocabulary size is manipulated when the HRN is applied to the 4-item recognition task. As was noted earlier, performance is virtually unaffected by vocabulary size until it reaches the number of hidden units. At this point there is a sharp drop. As shown by the learning curves, it takes longer for the HRN to reach asymptote as the number of items increases, but even when the vocabulary contains 16 items it is able to do so well within the 1500 epochs allowed.

3.5.5 Time Complexity

As mentioned in the previous section, the HRN takes longer to learn the recognition task as the number of vocabulary items increases. But, how quickly does the training time increase with the size of the input? Figure 3.12 shows the number of epochs the HRN required to reach its maximum performance plotted against the list length.

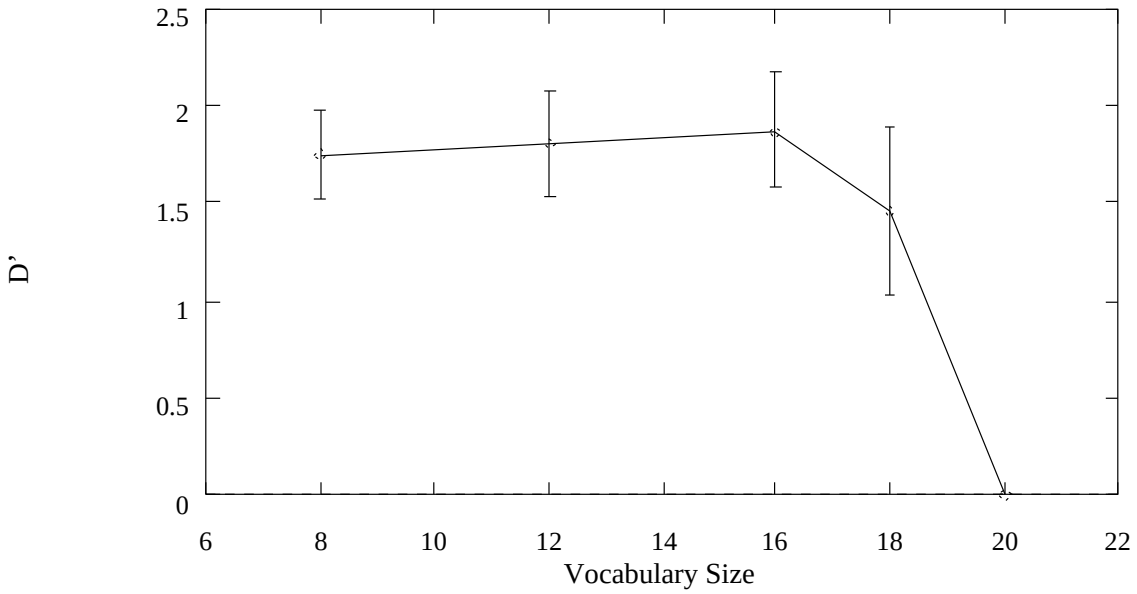

Figure 3.11: Performance (D') of the HRN as a function of the number of vocabulary items on the 4-item recognition task. The size of the vocabulary seems to make little difference until it reaches 20 when it drops to below chance. The bars indicate the 95% confidence intervals.

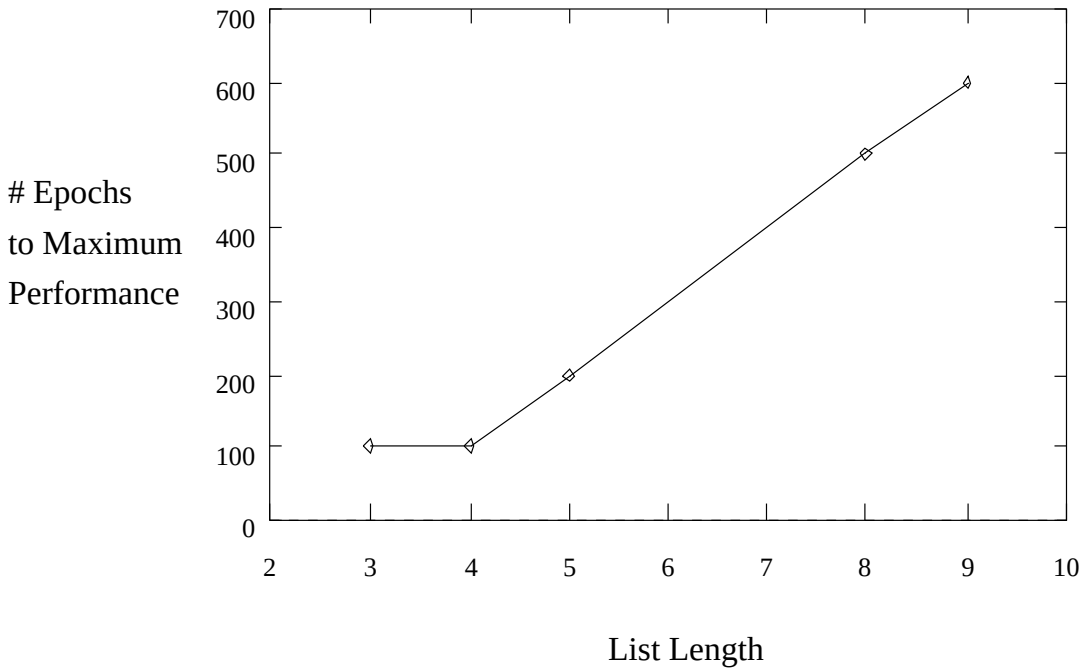

Figure 3.12: Time complexity of the HRN on the recognition task. The number of epochs required for the HRN to reach maximum performance versus the list length (i.e. size of the input). The growth is linear.

Although the result is not conclusive, because only small values of the list length have been sampled, the graph suggests that training time (in epochs) grows in an approximately linear fashion with the list length.

3.6 Hidden Unit Analysis

In order for the HRN to operate successfully it must form stable representations of the input items on its hidden units. Figure 3.13 shows the hidden patterns of the HRN plotted on the first two principal components[11]. Two hundred sequences each containing five patterns were presented and, hence, figure 3.13 shows 1000 hidden unit patterns. The dominant feature is the separation on the input pattern. To successfully autoassociate the input patterns the network maps all input/context patterns that correspond to the same input onto similar locations in hidden unit space, and attempts to separate these clusters from each other. Considering that there are on average 143 of each input pattern the clusters are very tight.

In contrast, position information is only poorly retained. Figure 3.14 shows the HRN hidden unit patterns plotted on the first two canonical discriminants[12] when grouped on position (i.e. 1, 2, 3, or P for probe). The probe and third position patterns are reasonably well separated, but the first and second position patterns show very little separation. Figure 3.15 shows the CDA plots of the hidden patterns grouped on position (i.e. 1, 2, 3, or P for probe) and on letter (i.e. a, b, c, d, e, or f) to the same scale. Clearly, the separation of the letters is much greater than the separation of the positions. Unlike the serial recall task explored by Wiles and Phillips (1991), position information is not required to do the recognition task and, hence, separation on the basis of position is poor[13].

In contrasting the SRN and HRN hidden unit representations, the first point to note is that the HRN is better at recognising patterns from the study list than the SRN. Examining the hidden patterns on the "Answer" timestep shows that the

[11] Koltz and Johnson (1982) provide a description of Principal Components Analysis (PCA) and Elman (1990) demonstrates its application to hidden unit analysis.

[12] Koltz and Johnson (1982) provide a description of Canonical Discriminants Analysis (CDA) and Wiles and Bloesch (1992) demonstrate its application to hidden unit analysis.

[13] If more hidden units are employed then one does see residual position information. The networks from which these plots were derived contained seven hidden units.

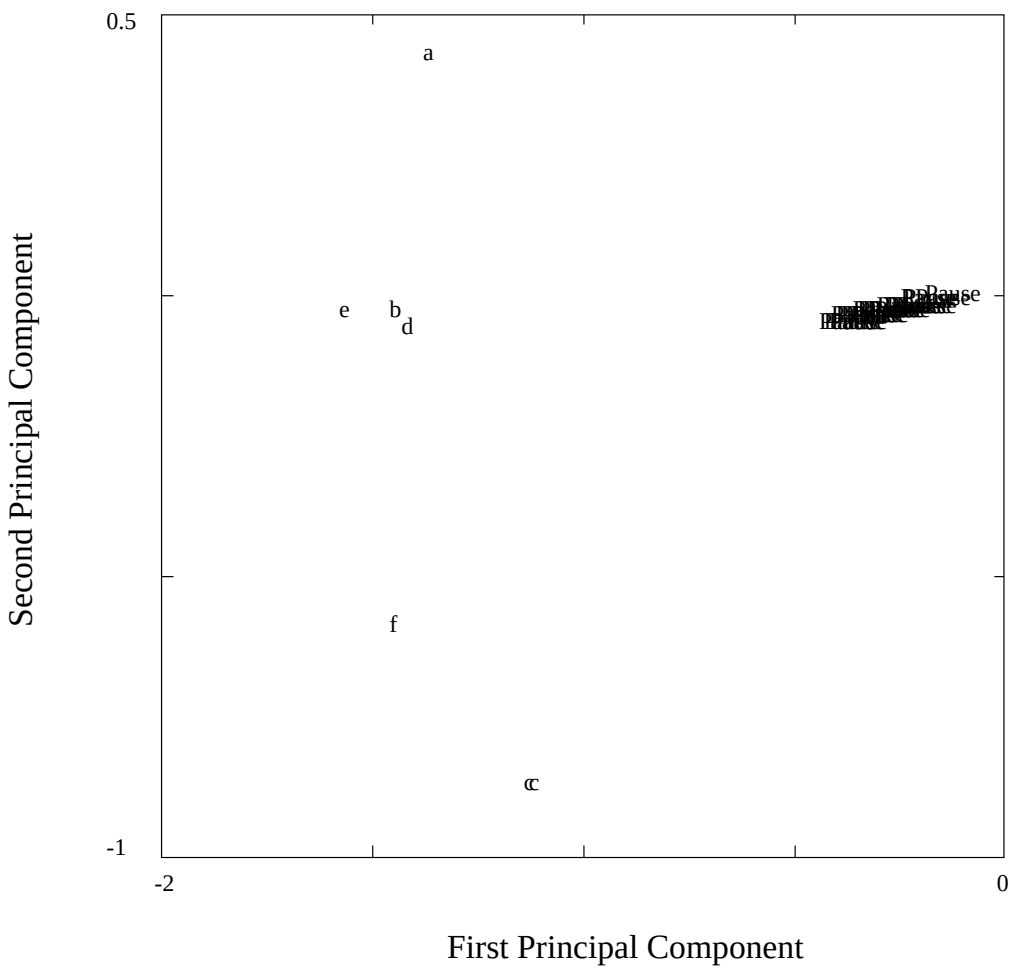

Figure 3.13: Principal components plot of the hidden patterns of the HRN. The HRN was run on 200 sequences and, hence, 1000 patterns are shown here. The hidden unit patterns cluster very tightly around the input pattern as a consequence of the autoassociation of the input.

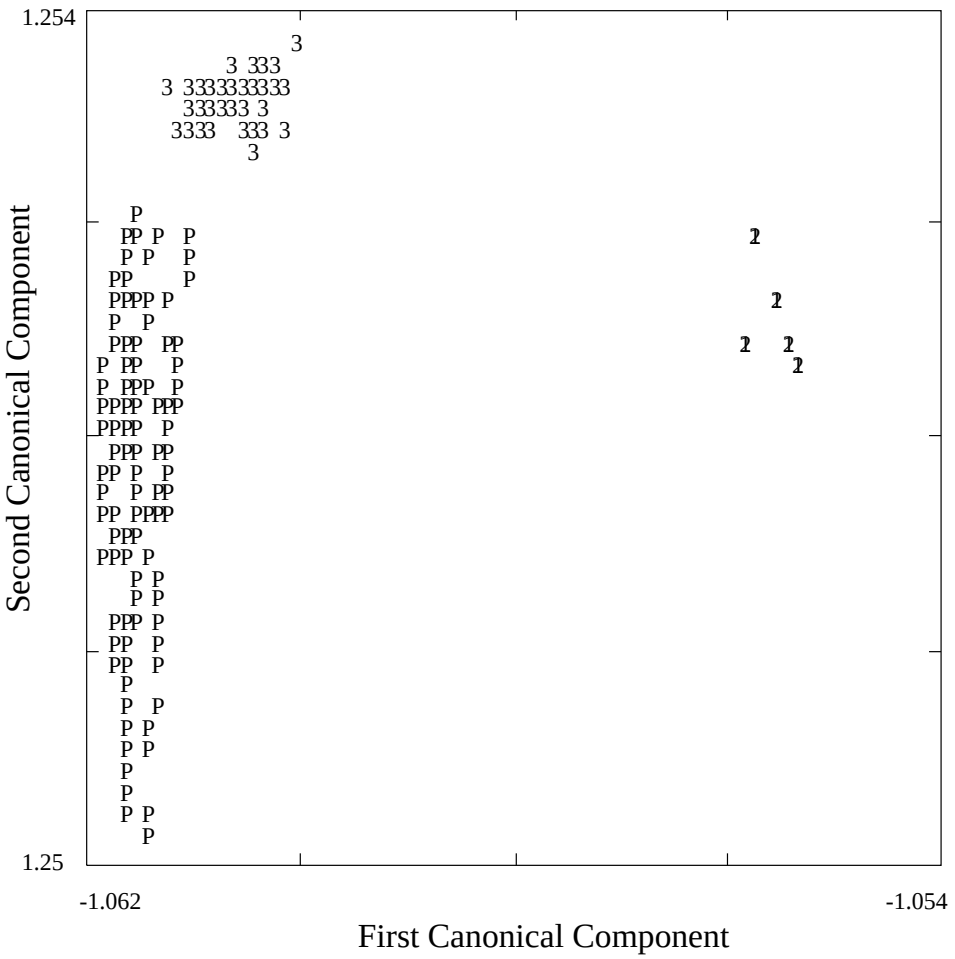

Figure 3.14: Canonical discriminants plot of the hidden patterns of the HRN grouped on position. The labels indicate first (i.e. 1), second (i.e. 2), third (i.e. 3) and probe (i.e. P) items. While the probe and third position patterns are reasonably well separated, the first and second positions show very little separation. Furthermore, the range of both axes is very small. Position information is retained, but poorly.

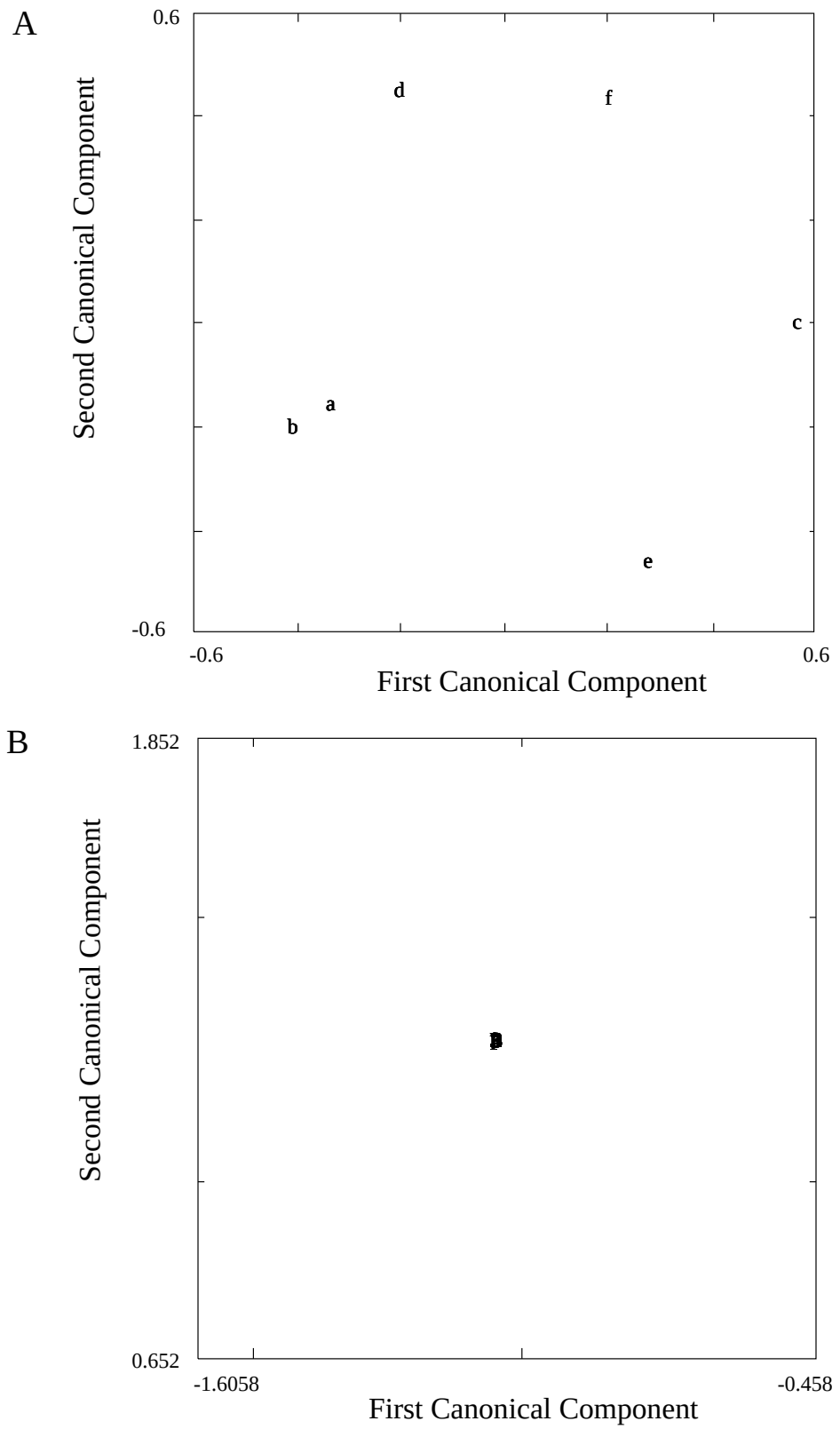

Figure 3.15: Canonical discriminants plots of the hidden patterns of the HRN grouped on input and on position. Figure A shows the CDA plot when the patterns are grouped on input (i.e. a, b, c, d, e, or f). Figure B shows in the same scale the CDA plot when the patterns are grouped on position (i.e. 1, 2, 3, or P for probe). Separation when grouped on letter is much greater than the separation when the hidden patterns are grouped on position. Such a pattern is expected since the recognition task does not require position information. In contrast, the letter information is required at every timestep.

Target and Distractor patterns are not as well separated in the SRN (see figure 3.16). It is more difficult to construct a decision boundary that will correctly classify these patterns as "Yes" or "No" responses.

While the SRN must form a representation of the entire list in its hidden unit activation patterns, the HRN can rely on the Hebbian memory to store items. Figures 3.17 and 3.18 are Hierarchical Cluster Diagrams (HCA, Elman, 1989) of the hidden patterns after the final study item has been input. At this stage, the SRN must have a representation of the entire list. In contrast, the HRN need only separate the current items. Items from the first and second positions of the list are stored in the memory. The labels in figures 3.17 and 3.18 indicate whether the corresponding study list contained an "a" and, if so, whether that 'a' occurred in the third position. In both the HRN and SRN the "a"s in the third position are well separated from the rest of the patterns. In the HRN, however, first and second position "a"s are mixed with the other patterns. By contrast, the SRN "a"s are well separated regardless of position. The learning task that the HRN must solve is easier than that of the SRN.

3.7 Conclusions

The introduction (section 3.1) outlined six criteria for a model of human memory, and in this section the HRN's performance is evaluated with respect to these criteria.

The first of the criteria was the learning of representation. The analysis of the hidden unit patterns of the HRN demonstrates that it is the case that the model learns a representational landscape. In the recognition task explored in this paper, all items were equal. The input patterns were orthogonal and all items fill identical functional roles. Consequently, the patterns were evenly distributed throughout the space. In tasks in which either the input patterns vary or the items are required to fill different functional roles (such as Elman's, 1989, prediction task) the hidden unit space would be more structured, and it would be possible to assess how the memory performs on perceptually and semantically related items.

The development of decision boundaries was the second criterion. In the HRN, the decisions are made by the backpropagation connections from the context-to-hidden layer and from the hidden-to-output layer. Hence, the HRN does satisfy the

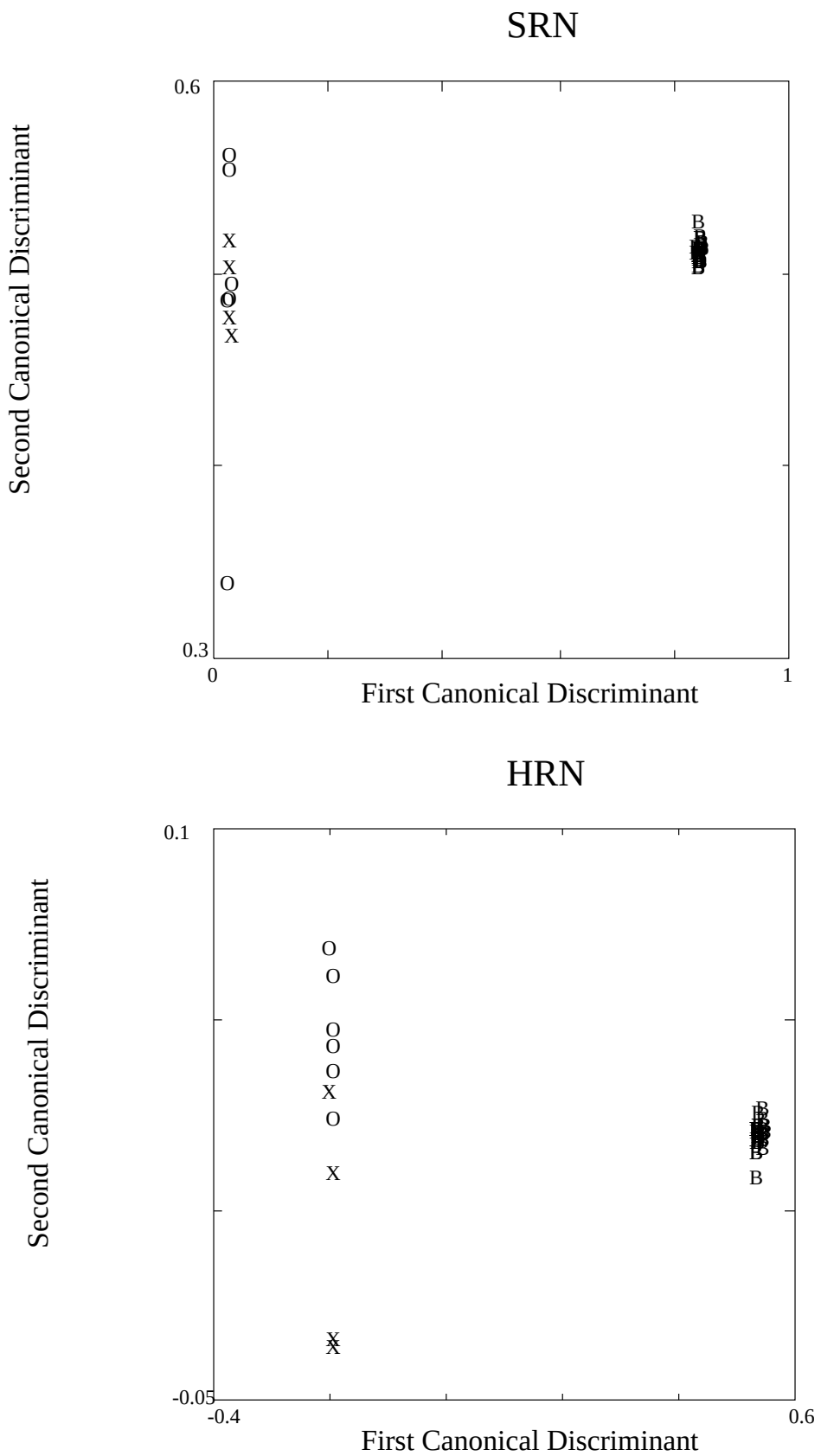

Figure 3.16: Canonical discriminant plots of the hidden unit patterns of the SRN and HRN grouped on output. The separation of target patterns, denoted by "O", and distractor patterns denoted by "X" is better for the HRN than for the SRN translating into improved performance for the HRN. The points labelled "B" correspond to "Blank" outputs.

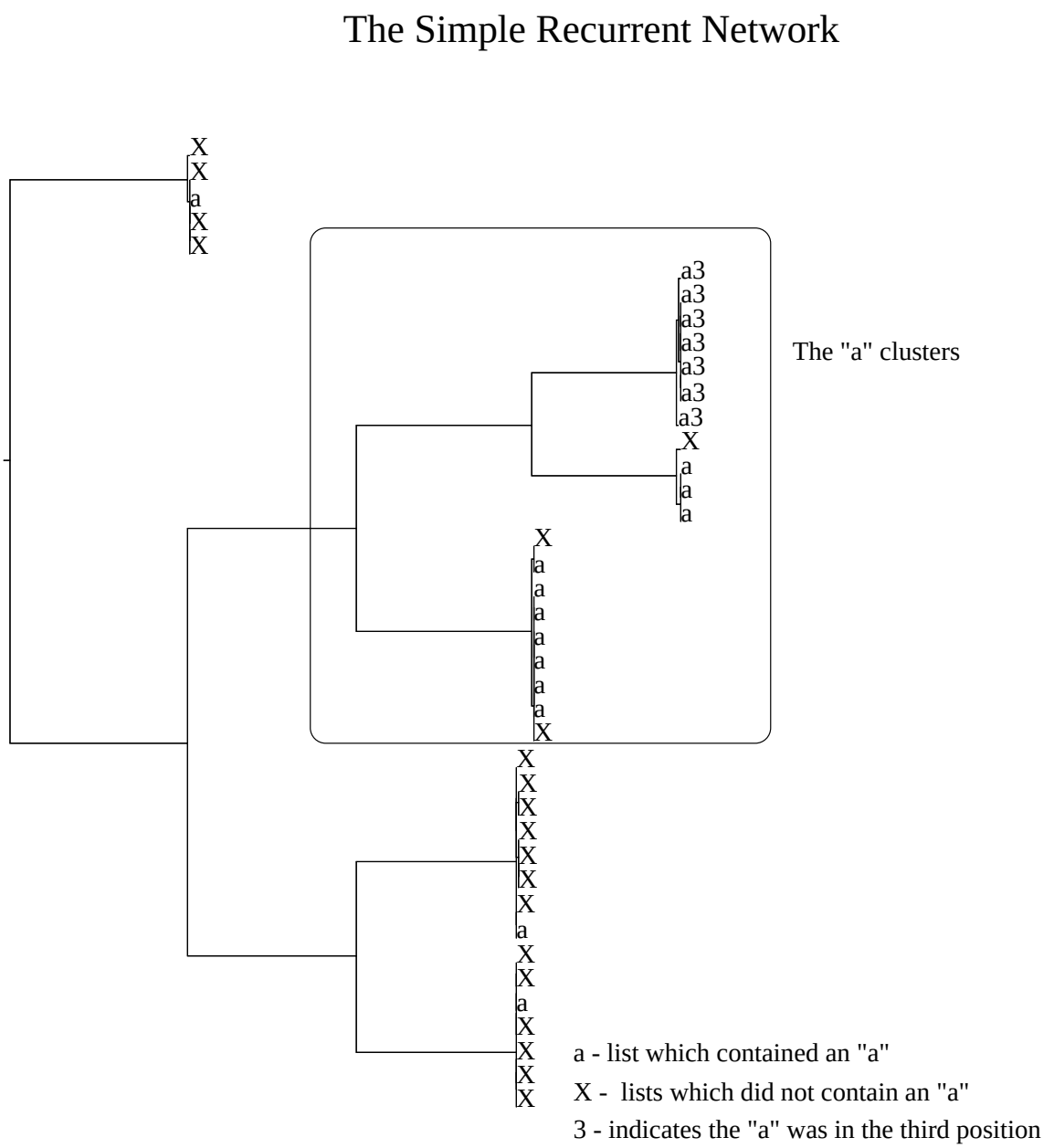

Figure 3.17: Hierarchical Cluster Analysis (HCA) of the hidden unit patterns of the SRN after the study list has been input. Note that the sequences that contain an "a" are cluster together. Such an organisation occurs since the SRN must respond with an "a" regardless of which position the "a" occurred in.

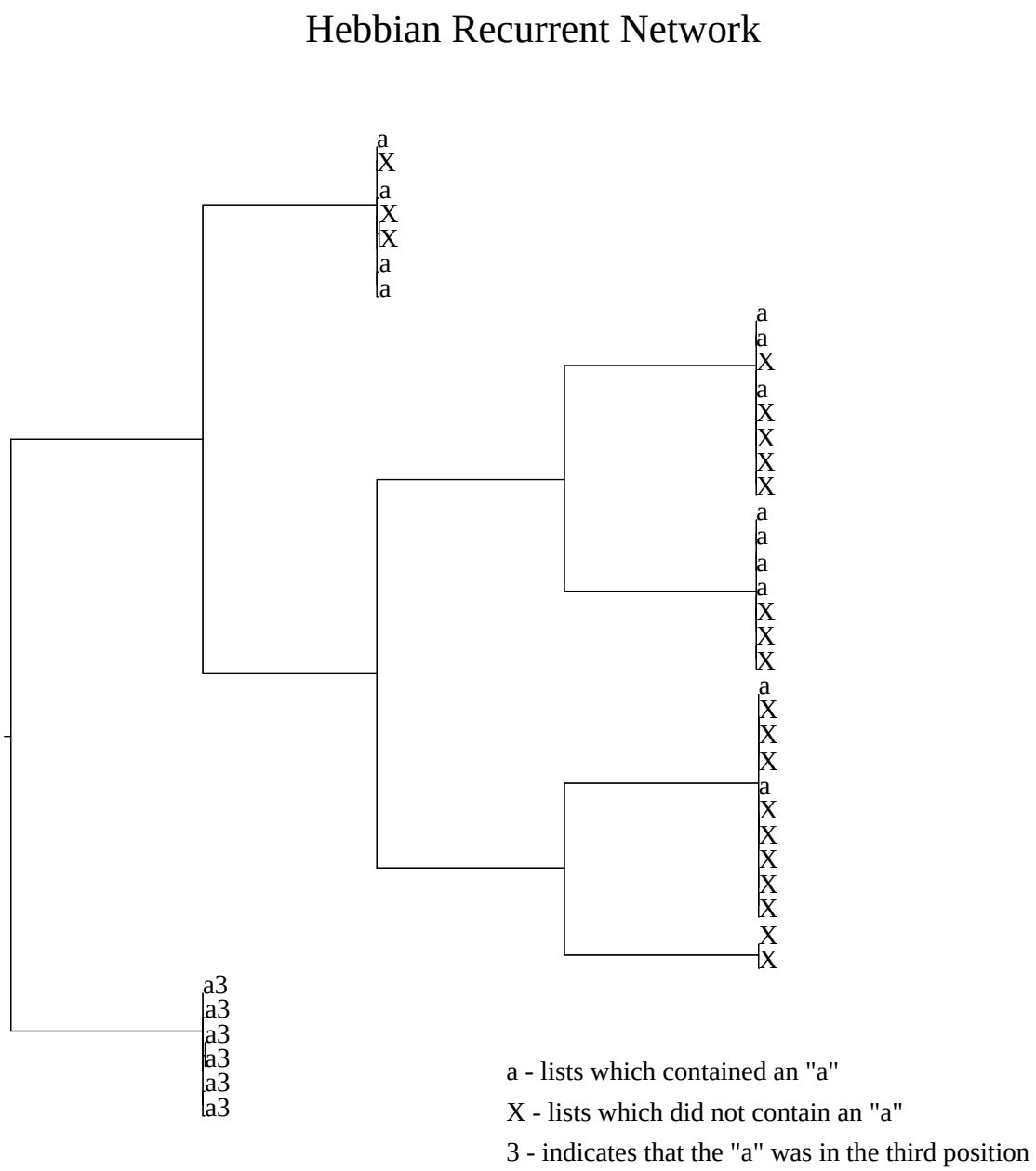

Figure 3.18: Hierarchical Cluster Analysis (HCA) of the hidden unit patterns of the HRN after the study list has been input. In contrast to the cluster analysis of the SRN, the HRN clusters only the third position "a"s well. The first and second position "a"s are mixed in with the non "a" patterns. The HRN does not need to separate these patterns in the hidden unit patterns since they are retained in the Hebbian memory.

condition of learning its decision criteria. Furthermore, as training progressed the performance increased, indicating that the decision criteria was being altered so as to better separate the target patterns from the distractors.

The third criterion was the development of control regimes. In such a simple task the degree of control required was limited. The only major distinction was between the study phase in which the network was required to output a "Blank" and the decision phase when the network responded with either "Yes" or "No" indicating whether it considered the probe item to have been in the study list. While it often took several hundred epochs of training before the network began to correctly determine whether to output "Yes" or "No", the control question of whether to output a "Blank" or a decision was typically learned very quickly, usually well within one hundred epochs. While this example is very simple it demonstrates that the HRN is capable of acquiring a control regime.

Learning is only useful if it can be done in a feasible time period. The time complexity of learning seems to be approximately linear in the length of the lists (and size of the vocabulary) with which the HRN is tested.

The first of the memory criteria was the degree of interference. The Hebbian memory of the HRN allowed it to retain items without significant interference. The HRN inherits the performance characteristics from the matrix memory and maintains its performance (at least until vocabulary size becomes larger than the number of hidden units) with extra items decrementing performance only marginally.

Since human vocabularies are in the order of 10000 items and the number of items that could conceivably be remembered is even greater, it is critical that a model of memory be able to store and retrieve a large number of items. The ability of the HRN to generalise well even as the number of vocabulary items increased is of particular importance, and is one of the major distinguishing factors between it and the SRN.

The last of the criteria was the rapid binding in memory of already established representations. The possible representations are developed by the backpropagation mechanism over the course of training. Specific memories, however, are stored in the Hebbian weights. Hence, memories are laid in a single timestep, while representations are formed over a much longer timespan. It is the dual memory architecture

that avoids catastrophic interference, allows for the significant improvement in generalisation, and accounts for the dramatically different timespans of memory and learning.

While there is still work to be done to establish the extent of the capabilities of the HRN, it has been demonstrated that by integrating insights from backpropagation models with those from the mathematical memory modelling literature it is possible to fulfill both learning and memory criteria.

Chapter 4

Environment Meets Mechanism: The Word Frequency Example

4.1 Introduction

Chapter two demonstrated how one might go about quantifying the statistics of the environment. Chapter three developed a *learning* mechanism (i.e. the Hebbian Recurrent Network) that retained some of the essential characteristics of a model of memory. This chapter completes the interactive story by bringing the environmental findings to bear on the HRN to produce a model of human performance that accounts for the word frequency effect in recognition.

To investigate the affect of word frequency in recognition, subjects are presented with a list of study words. The list is constructed such that some of the words occur with high frequency in the language and others occur with low frequency. The subjects are then presented with a second list of words some of which occurred in the study list and some which did not. They are required to determine if each of the words occurred in the study list. Typically the subjects are more accurate on the low frequency words than on the high frequency words (Allen & Garton, 1968; Gorman, 1961; Kinsbourne & George, 1974; McCormack & Swenson, 1972; Schulman, 1967; Schulman & Lovelace, 1970; Shepard, 1967; Underwood & Freund, 1970).

Chapter two outlines the difficulty that arises when current accounts of the word frequency effect are applied to data concerning the impact of word knowledge on

recognition (Allen & Garton, 1968). Contrary to current predictions performance seems to improve if subjects are familiar with the low frequency words. It was suggested that an environmental explanation in which a parallel was drawn between episodic recognition and recurrences of a word in a given context could account for both the general word frequency effect and the word knowledge data. To this end, an environmental study was conducted that provided evidence that low frequency words were on average more likely to recur within a given context than high frequency words. In this chapter, this environmental information is used to construct a training set for a the HRN, providing a processing mechanism for the effect.

In an interactive account of the word frequency effect in recognition, the important issue is how the environment and the mechanism interact to produce behaviour. *A priori* the HRN does not distinguish between high and low frequency words. As in the simulations presented in chapter three, words are represented by orthogonal vectors and there is no oracle to classify words into high and low frequency groups. No assumption of the mechanism, then, can account for any difference in performance. The strategy is to construct a training set that embodies the statistical premises that were verified in the environmental analysis. As the HRN is trained, it must learn to respond more accurately to low frequency words. Any performance differences will come about as a consequence of the interaction of the mechanism and its environmentally constrained training set.

The limitations of simulation on large vocabularies preclude any quantitative predictions. Hence, the aim is not to produce a quantitative model of the word frequency effect. The aim is to demonstrate how a model that makes no assumptions about the input materials can come to display qualitative behaviour similar to human subjects when exposed to the appropriate environment. The interest is primarily in the interactivist principle.

To start, the environmental analysis of the word frequency effect from chapter two will be reviewed. The next two sections outline the method and the results of the simulations. The chapter ends with a discussion of this application of interactivism.

4.2 The Environmental Context

In the environmental analysis of the word frequency effect conducted in chapter two, empirical evidence was presented that suggested that low frequency words are more likely than high frequency words to recur within a context. Low frequency words seem to be more context specific. If they have occurred on one occasion they are quite likely to occur again (see chapter two, figure 2.4). That is, the conditional probability of a word occurring given that it has already occurred in this context is higher for low frequency words:

$$P(H_i \mid H_i \in PreviousContext) < P(L_j \mid L_j \in PreviousContext) \tag{4.1}$$

where H_i is a high frequency word, L_j is a low frequency word and $Previous\ Context$ contains all words in the current context that precede the current word[1].

In addition to this constraint, low frequency words occur less frequently in general than high frequency words. Hence:

$$P(H_i) > P(L_j) \tag{4.2}$$

While there are many other statistics embodied in everyday prose, these two affect the probabilities of the occurrence of items as a consequence of frequency and, consequently, are the constraints that will precipitate any performance difference between high and low frequency words in the simulations.

4.3 Simulation 1: Word frequency and the HRN

The aim of the first simulation was to test the relative performance of high and low frequency words when the HRN is trained on a data set constrained by the environmental statistics. It was hypothesised that since low frequency words have a higher recurrence rate, they would be more accurately recognised than high frequency words despite the fact that in the training set high frequency words occur

[1]As an aside, it was also proposed in chapter two that the probabilities of recurrence of low frequency words decrease more quickly than their high frequency counterparts. Given that low frequency words by definition occur less frequently than high frequency words the probability of a high frequency word must exceed that of a low frequency word at some stage. Just after the occurrence of the word the probability of the low frequency word is higher, hence, the occurrence probability of a low frequency word must on average decrease more quickly than that of a high frequency word (see chapter two, figure 2.5).

with greater probability than low frequency words during both the study and test phases. In the simulations that follow, orthogonal vectors have been used to represent the items. It is likely that some of the other variables that are correlated with word frequency (e.g. orthographic distinctiveness) would result in an input space in which high and low frequency words were separated on some dimension. In these simulations, however, the objective was to demonstrate that recurrence rate information alone was sufficient to produce the effect.

4.3.1 Method

Network Architecture and Parameters

Figure 4.1 shows the architecture that was used. All items were encoded locally at input and output. There were 16 input units, one for each of the words used and an extra "Pause" input that indicated that there was no input at this timestep. The non zero value of each vector was set to 0.7 rather than 1 so that the targets were reachable (Touretzky & Pomerleau, 1991). For example the vector representing the third high frequency word was (0 0 0 0.7 0 0 0 0 0 0 0 0 0 0 0 0). Sixteen hidden units (and sixteen context units) and nineteen output units were used. Three of the output units corresponded to "Blank" (i.e. do not output anything at this stage), "Yes" and "No" (i.e. was the test item in the study list or not?). The other 16 outputs were the current inputs. As in the previous chapter, the inputs were autoassociated to ensure separation of the hidden unit patterns corresponding to different input patterns.

The backpropagation part of the network was trained with a learning rate of 0.05 and the memory decay rate was set to 0.2. Initial weights were selected from a uniform distribution between -1 and 1 and then updated after each pattern presentation. The simulations were run for 1500 epochs by which time the performance on the simulations had reached asymptote.

Training Set Construction

As in the previous chapter, the episodic recognition task was simplified to presenting a list of items followed by a single test item. A "pause" pattern was then input and it was at this stage that the network was expected to respond (i.e. "Yes" or "No"

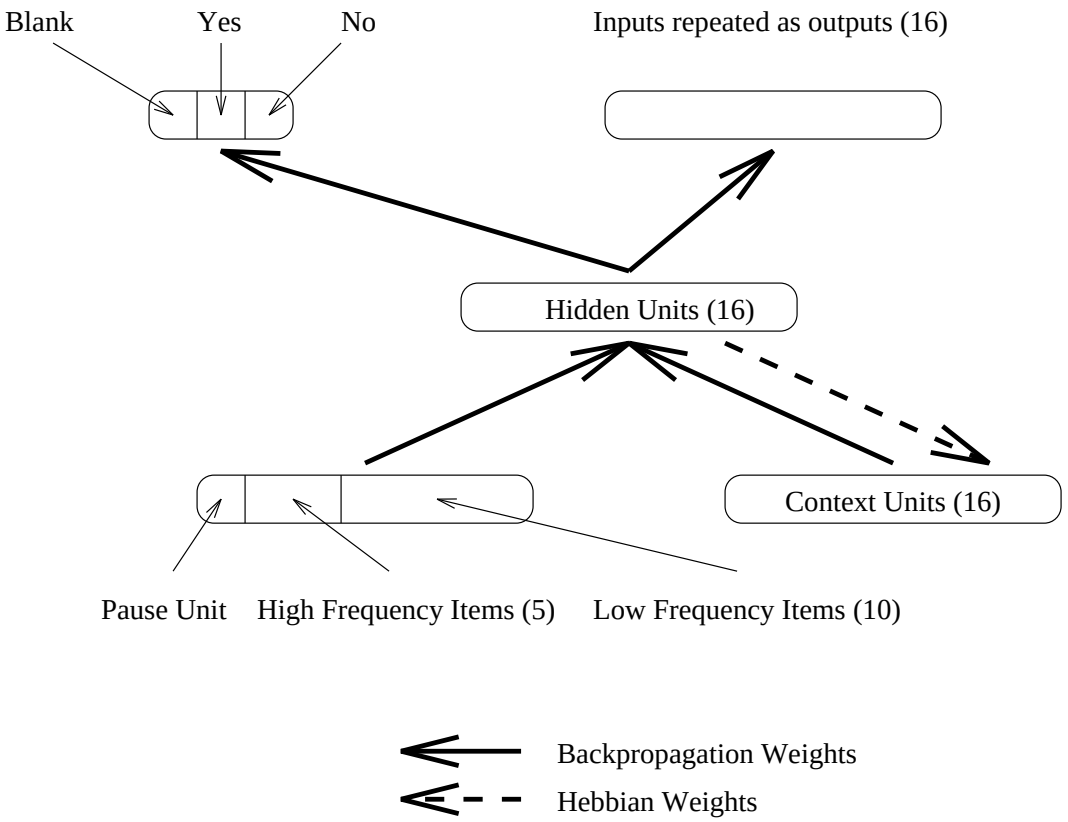

Figure 4.1: The HRN architecture for the word frequency simulations. The architecture is a Hebbian Recurrent Network (HRN) as outlined in chapter three. The inputs and context units are connected to the hidden layer by backpropagation weights as is the hidden layer to the outputs. The hidden layer is connected to the context layer by Hebbian weights. The symmetric sigmoid (i.e., $\tanh(\frac{x}{2})$) was used throughout. The input vectors consisted of sixteen components. In each vector all except one of the components was zero. The final component was set to 0.7 and coded the input item that was being presented. Five of the units coded for high frequency words, ten for low frequency words and one was a "pause" signal that indicated that no input was being presented at that time. The "pause" input was presented when the decision whether the test item was in the study list was being made. Sixteen of the outputs were the inputs repeated. Requiring the inputs as outputs serves to separate the input patterns (see chapter three). The other three outputs coded for "blank", "yes" and "no". Either "yes" or"no" was expected when the recognition decision was made and the "blank" pattern was expected at all other times.

depending on whether the item was in the list or not).

To carry out the simulations with the available time and computational resources, it was necessary to limit the size of the training set by restricting both the number of items in each list and the vocabulary from which these items were chosen. Study list size and vocabulary size trade off. In order to ensure that each of the items was presented as a test item with a variety of study lists (both in which it occurred and in which it was absent), it was necessary to present a large number of sequences. To ensure stability of the results several replications (simulating several subjects) were required. Training sets consisting of 500 sequences with three item study lists chosen from a vocabulary of fifteen (a training set of 2500 patterns) were constructed representing a compromise of these conflicting demands.

The fifteen vocabulary items where then divided into high and low frequency words. To reflect the fact that there are more low frequency words than high frequency words (c.f. Francis & Kucera, 1982), five of the vocabulary items were considered to be high frequency words and ten were designated low frequency words.

Each sequence consisted of three study items and a test[2] item. When constructing the study list, items were sampled with replacement and high frequency words were sampled at twice the rate of the low frequency words.

Once the study list had been finalised it was possible to select test items that did occur in the study list (i.e. targets) or items that did not occur (i.e. distractors). In addition, it was possible to select test items that were either high frequency or low frequency. The probabilities of each of these conditions must be carefully selected so that the environmental constraints are met. The chosen values are presented in table 4.1. So, for instance, high frequency words from the study list were chosen as test words with a probability of 0.1.

To establish that the training set does satisfy the environmental constraints, the constraints are re-expressed in terms of the training set items. Then, the training set is analysed to ensure that it successfully embodied the environmental statistics.

[2]The word "test" is used to refer to both the test phase in the study-test paradigm used for the memory experiment and to refer to the test phase of the network as opposed to the training phase. Hence, each sequence in the training and test sets contains a study and test phase. This terminology is inherited from the neural networks and memory fields.

Table 4.1: Training set statistics for the word frequency simulation. The table shows the probabilities with which different conditions occurred in the training set.

Condition	Probability
High Frequency Target	.1
Low Frequency Target	.35
High Frequency Distractor	.25
Low Frequency Distractor	.3

Re-expressing the environmental statistics

It is assumed that the study items should reflect the statistics of the previous context and the test item should correspond to the current item in inequality 4.1.

When the substitutions are made, inequality 4.1 becomes:

$$P(H_i \in Test \mid H_i \in StudyLst) < P(L_j \in Test \mid L_j \in StudyLst) \tag{4.3}$$

where H_i is a high frequency item, L_j is a low frequency item, $Test$ is the singleton set containing the test item and $StudyLst$ contains the items that were presented during the study phase.

Inequality 4.2 should apply to both the study list and the test list. Hence, the following constraints also apply to the training set.

$$P(H_i \in StudyLst) > P(L_j \in StudyLst) \tag{4.4}$$

$$P(H_i \in Test) > P(L_j \in Test) \tag{4.5}$$

Checking the training set

When constructing the study list, high frequency items were chosen with twice the frequency of low frequency items. Hence, the probability that H_i occurs in a given position is twice the probability that L_j occurs in a given position.

$$P(H_i) = 2P(L_j) \tag{4.6}$$

Since there are five high frequency words and ten low frequency words:

$$1 = 5P(H_i) + 10P(L_j) \tag{4.7}$$

Hence, the probability of a given low frequency word occurring in a given position is:

$$P(L_j) = 1/20 \tag{4.8}$$

and the probability that a given high frequency word occurs in a given position is:

$$P(H_i) = 1/10 \tag{4.9}$$

Now the probability that a given word occurred in the study list is one minus the probability that it didn't occur in any of the three positions. Hence,

$$P(L_j \in StudyLst) = 1 - (19/20)^3 = .14 \tag{4.10}$$

and

$$P(H_i \in StudyLst) = 1 - (9/10)^3 = .27 \tag{4.11}$$

Consequently, high frequency words occur with greater frequency than low frequency words within the study list (i.e., inequality 4.4 is satisfied).

Now consider the recurrence inequality 4.3. Since there are five high frequency words, the probability of a high frequency target is five times the probability of a given high frequency target:

$$P(HighFrequencyTarget) = 5 * P(H_i \in Test \ \& \ H_i \in StudyLst) \tag{4.12}$$

hence,

$$P(H_i \in Test \,|\, H_i \in StudyLst) = \frac{P(H_i \in Test \,\&\, H_i \in StudyLst)}{P(H_i \in StudyLst)} = .07 \tag{4.13}$$

Similarly there are ten low frequency words, hence:

$$P(LowFrequencyTarget) = 10 * P(L_i \in Test \ \& \ L_i \in StudyLst) \tag{4.14}$$

and

$$P(L_i \in Test \mid L_i \in StudyLst) = \frac{P(L_i \in Test \ \& \ L_i \in StudyLst)}{P(L_i \in StudyLst)} = .24 \tag{4.15}$$

The probability that a low frequency word in the study list is a test item is greater than the probability of its high frequency counterpart (i.e., inequality 4.3 holds).

Finally, when examining the relative frequencies of high and low frequency words as test items (i.e., inequality 4.5):

$$P(H_i \in Test) = \frac{P(HighFreqTarget) + P(HighFreqDistract)}{NumberofHighFrequencyWords} = .07 \quad (4.16)$$

while,

$$P(L_i \in Test) = \frac{P(LowFreqTarget) + P(LowFreqDistract)}{NumberofLowFrequencyWords} = .065 \quad (4.17)$$

and inequality 4.5 is also satisfied.

High and low frequency test probabilities: The size of the difference between the probability of a high frequency test item and a low frequency test item is much smaller in the simulations than is the case in a normal text corpus. The difficulty arises in trying to satisfy all three conditions simultaneously. Lowering the probability of a low frequency word or increasing the probability of a high frequency word has the negative affect of decreasing the difference between $P(H_i \in Test \mid H_i \in StudyList)$ and $P(L_i \in Test \mid L_i \in StudyLst)$. With human size vocabularies this presents no problem, however, in order to make simulation feasible the vocabulary of the network must be restricted, making it impossible to maintain the quantitative statistics of everyday text. Therefore, as mentioned above, no quantitative claims can be made.

Procedure

Twenty simulations were conducted. Each simulation was taken to correspond to a different subject. Initial weights, and training and test sets (containing 500 and 200 sequences respectively) were generated for each simulation.

4.3.2 Results

For each network the *d'* values for high and low frequency words were calculated from their hit and false alarm rates as indicated in chapter three. Figure 4.2 shows the learning curve, that is, the performance measured by *d'* as a function of the training epoch. An analysis of variance with repeated measures on the frequency factor was conducted at 100, 200 and 1500 epochs. The performance of the low frequency words is significantly higher than that of the high frequency words in the first two hundred

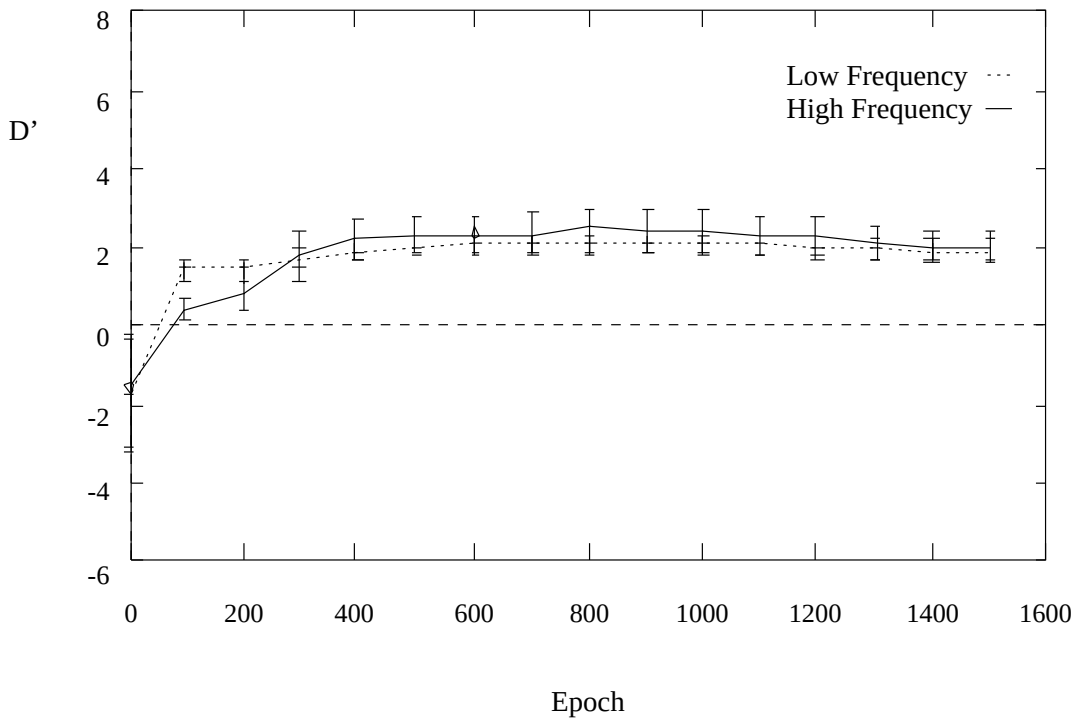

Figure 4.2: Training epoch verses performance (D') of the HRN on high and low frequency words. In the first two hundred epochs performance on the low frequency words exceeds that on the high frequency words as is the case with human subjects. As training progresses, however, the performance on the high frequency words improves more quickly, and in the limit there is no significant difference between the performance on the high and low frequency words. The bars represent the 95% confidence intervals.

epochs (100 epochs: $F(1,19) = 24.3$, $p < 0.0001$, 200 epochs: $F(1,19) = 17.027$, $p < 0.001$). However, the high frequency words catch up as training progresses until at the end of training there is no significant difference between the performance of the high and low frequency words, $F(1,19) = 0.036$, $p = 0.851$.

4.3.3 Discussion

At the beginning of training the performance is better for low frequency words. Since low frequency words are more frequently targets in the training set, the boundaries for the Yes/No decision at test adjust more quickly. Once the boundaries for the low frequency words are nearing their optimal position (relative to the training set) they change little as there is negligible error generated to drive an adjustment of the weights. As the training progresses the majority of the error is attributed to the high frequency boundaries. Hence, they change more quickly until they are equivalent in accuracy with the boundaries of low frequency words and the high

frequency performance rises to match that of the low frequency words. The network does not model human performance.

To address the difference between the performance of the model and human performance, it is useful to characterise the behaviour of the model in terms of the number of effective parameters. When using backpropagation with a multilayer architecture of sigmoid units, generalisation performance can often be improved by stopping training before the training set error has reached its minimum (Hinton, 1993). This occurs because when the weights are small the hidden unit activations come largely from the linear portion of the sigmoid. The network is equivalent to a single layer network (since the pre and post hidden layer matrices can be multiplied to give a single matrix). As training progresses, however, and the weights become larger, the range of activations extends into the nonlinear portion of the sigmoid. The single layer approximation becomes less accurate and the effective number of parameters increases.

At the start of training the performance on the low frequency words was greater than that on the high frequency words. If training was discontinued at 100 epochs the low frequency words would out perform high frequency words as is observed in human subjects. However, it would have to be assumed that the performance that is observed in human subjects is a consequence of the specific number of examples to which they have been exposed. It seems more reasonable to assume that the performance in adult subjects is close to asymptotic. The difficulty for the HRN as a model of relative recognition performance of high and low frequency words arises as the training progresses and the number of effective parameters increases.

What is required is a method for permanently decreasing the number of effective parameters so that the asymptotic performance will be the same as that at 100 epochs. In psychological terms, one might argue that the capacity of the system must be decreased.

A common method for decreasing the number of parameters is to reduce the number of hidden units. However, as was demonstrated in the previous chapter, performance in the HRN drops to chance if the number of hidden units falls below the number of inputs, since it is no longer possible to make the hidden patterns orthogonal and, hence, use the hebbian weights efficiently.

Another way of decreasing the number of effective parameters is to introduce a weight decay parameter (Moody, 1992). Each weight is reduced (or in the case of negative weights increased) by a small proportion at each pattern presentation. Weight decay encourages the activations to remain in the linear part of the sigmoid and, hence, the model complexity is reduced. With an appropriate weight decay the qualitative performance observed at 100 epochs could be maintained.

4.4 Simulation 2: Adding Weight Decay

Simulation two explored the affect of a weight decay parameter on asymptotic performance of the HRN. The aim was to reduce the number of effective parameters so that the high and low frequency words would compete and the performance seen in the first two hundred epochs would be maintained.

4.4.1 Method

The parameters and method used in the second simulation were identical to the first, except that the weight decay parameter was varied between 0 and 0.00005 (per pattern presentation). Twenty simulations were conducted at each value of the weight decay parameter.

4.4.2 Results

Figure 4.3 shows the learning curves for the recognition task for weight decay parameters between 0 and 0.00005.

The first point to notice is that performance improves dramatically with an appropriate weight decay. When the weight decay is at 0, performance reaches an average d' of 1.97 at 1500 epochs. When the weight decay is increased to 0.00001, the average performance jumps to 7.103. As the weight decay is increased beyond 0.00001, however, performance drops until it reaches zero when the weight decay is 0.00005.

Secondly, when the weight decay was 0.00002, 0.00003 or 0.00004 (figure 4.3 parts b, c, and d), the asymptotic performance of the low frequency words was significantly higher than that of the high frequency words (i.e., At 1500 epochs for a weight decay

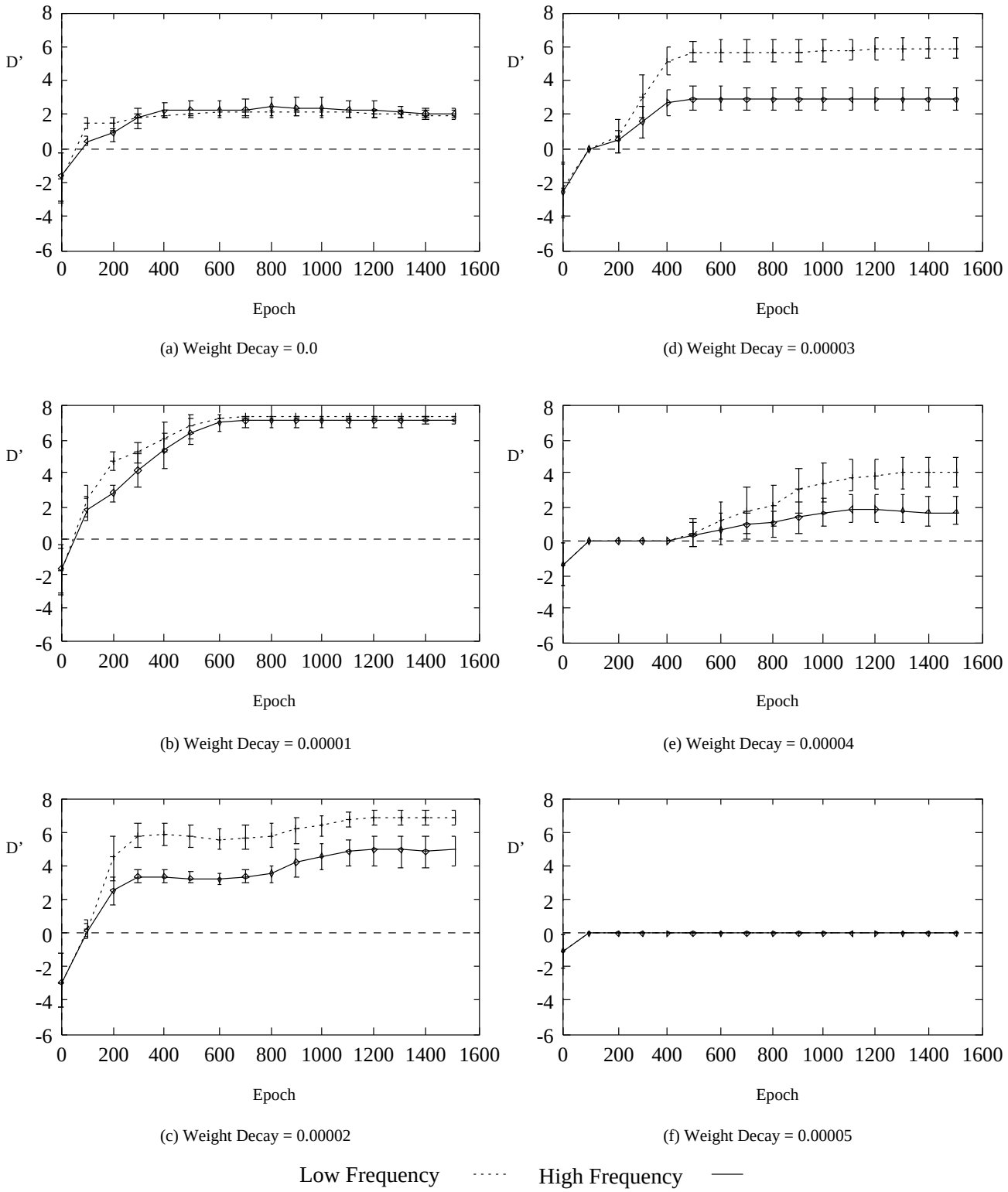

Figure 4.3: Learning curves for the HRN applied to the recognition task for several values of the weight decay parameter. There are two things to notice. Firstly, the average performance increases when a weight decay of 0.00001 is introduced and, then, decreases as the value is increased. Secondly, for the weight decays of 0.00002, 0.00003, and 0.00004 the performance on the low frequency words remains significantly higher than that for the high frequency words.

of 0.000002, F(1,19) = 33.005, p < 0.0001; for a weight decay of 0.000003, F(1,19) = 202.605, p < 0.0001; and for a weight decay of 0.000004, F(1,19) = 60.908, p < 0.0001).

4.4.3 Discussion

Introducing a weight decay decreases the number of effective parameters that the network can employ. When there are ample effective parameters to solve the task, increasing the weight decay improves the generalisation performance. As the weight decay is increased further, however, the number of effective parameters drops below that needed to solve the task and the performance drops again. There exists a weight decay, $\lambda_{optimal} > 0$ (in this case approximately 0.00001), that minimises the error on the test set (Moody, 1992).

The effective number of parameters is also important when considering the relative performance of the low and high frequency words. When there is no weight decay, as was the case in the previous simulation (reproduced in figure 4.3 part a), the performance of the high frequency words reaches the same levels as that of the low frequency words.

The weight decay can be adjusted, however, so that the low frequency and high frequency words must compete for effective parameters (see figure 4.3, part d). Since the low frequency words are targets more frequently, their accuracy is improved at the expense of the high frequency words. Hence, the low frequency performance asymptotes at a higher level than the high frequency performance, as is the case with human subjects. The number of effective parameters might be thought of as a form of capacity that is allotted to high and low frequency words on the basis of their recurrence probabilities.

Clearly, the accuracy displayed by the network exceeds that found in the laboratory. There are a number of ways in which the *d'* values could have been reduced. The weight decay parameter could have been optimised or a constant noise source could have been introduced. However, neither of these additions would have added substantially to the theoretical content of the model.

Of more interest is the nature of the hidden representations that are formed by the HRN. To demonstrate that the low frequency advantage could be generated by

the recurrence statistics alone, the input representations and the functional roles of the items were made identical. Hence, the hidden unit representations that were formed showed no separation of the high and low frequency word categories. However, there are many variables correlated with frequency that have been identified in the literature and these variables may form the basis for separation. For instance, on the basis of orthographic distinctiveness one would expect the input representations of high and low frequency words to differ systematically. Likewise, correlated variables such as number of associates suggest that high and low frequency words differ in terms of the functional roles that they assume. Both of these factors would lead to a separation of the hidden unit patterns and would allow the model to account for paradigms such as release from proactive interference as frequency category is changed.

4.5 Conclusions

The purpose of these simulations was to demonstrate the plausibility of an interactive model of the word frequency effect and to provide an example of how one might construct interactive accounts of memory phenomena. The critical ingredients of an interactivist explanation are an empirical account of the statistics of the environment and a mechanism that is sensitive to these statistics.

The first ingredient was addressed in chapter two, where it was discovered that low frequency words have a greater recurrence probability than high frequency words within a context. The second ingredient was the backpropagation based HRN architecture, which was discussed in chapter three. The task of this chapter has been to apply the HRN to the recognition task, using a training set that reflected the statistics of the environment, and to assess whether the HRN is sufficiently sensitive to acquire the appropriate performance.

The results demonstrate that the HRN was able to produce a low frequency word advantage in the recognition task on which it was trained despite the fact that prior to training the architecture makes no distinction between high and low frequency words. Unlike many current models[3], the processes of which are fixed, the HRN

[3]Such as SAM (Raaijmakers & Shiffrin, 1981), TODAM (Murdock, 1982), CHARM (Eich, 1982), Minerva II (Hintzman, 1984) or the matrix model (Pike, 1984).

relies not only on a mechanism but also on the subtle statistics that are embedded in the environment with which the memory system interacts.

Furthermore, if one assumes that the experience of experts would involve higher recurrence rates for words within their domains, the model can account for a low frequency advantage on general words and explain the improvement in accuracy of subjects with prior experience of a set of words (Allen & Garton, 1968), a finding that is problematic for most other accounts.

Chapter 5

Environment meets Mechanism: The List Strength Example

5.1 Introduction

The aim of this chapter is to provide an interactive explanation of the null list strength effect in recognition memory (Ratcliff et al., 1990; Murnane & Shiffrin, 1991b; Yonelinas et al., 1992). In chapter two, the statistics relevant to the effect were investigated and it was postulated that in conjunction with an appropriate optimising memory system, they could account for the phenomenon. In this chapter, the Hebbian Recurrent Network (HRN) is proposed as such a memory system. The HRN is trained with a data set constrained by the environmental statistics and its performance is then compared against that found in the experimental setting.

The chapter is organised in a similar fashion to the word frequency chapter (chapter four). It begins by reviewing the environmental evidence relevant to the list strength effect both from Anderson and Schooler (1991) and from the empirical work in the environment chapter (chapter two). Then there is a description of the simulation method and results. Finally, the discussion and conclusions contrast the interactive explanation of the null list strength effect provided here with current mechanistic accounts.

5.2 The Environmental Context

The purpose of this section is to review the constraints from the environment that must be captured by the training set.

Most models of recognition memory can be extended to model the null list strength effect. In doing so, however, they often loose the ability to account for either the main effect of strength (i.e., the item strength effect) or the affect of length (Shiffrin et al., 1990). For instance, Murdock and Kahana (1992) propose that TODAM (Murdock, 1982) be extended by including noise from pre-study list items. Any affect of list strength is masked by the additional noise, and a small positive list strength effect is predicted. As Chappell (1993) points out, however, the additional noise will have the same affect on items in a length condition. Hence, the noise will also swamp any list length effect. Similarly, in an attempt to rescue the current global matching models (see Humphreys et al. (1989) for a review) one might propose that neither repeating items nor presenting them for a longer duration affects the strength of these items in memory. Again this would be problematic because there would be no basis for the main effect of strength.

Hence, when modelling the null list strength effect in recognition one must simultaneously model both the length and strength effects. In the next three sections, then, the environmental evidence relevant to each of these effects is reviewed.

The Item Strength Effect: The analysis of Anderson and Schooler (1991) showed that the recurrence probability of an item is a linear function of the frequency of that item. The constant of proportionality ranges between 0.76 and 1.0 depending on the type of material studied (see figure 2.7).

In the environment chapter (chapter two) it was suggested that this analysis could be improved by removing the component of the prediction due to the frequency of a word in general. In this way the environmental analysis matches the experimental situation more closely, in that the strength of items is manipulated independently of the general frequency of the word. When the general word frequency was factored out however, the contextual frequency of occurrence was still a strong positive predictor of recurrence probability. That is, the probability that a word would occur was greater if it had occurred many times in the preceding context

than if it had occurred infrequently:

$$P(W_i \mid W_i \; is \; strong) > P(W_j \mid W_j \; is \; weak) \tag{5.1}$$

where W_i is a word and W_i *is strong* means that W_i occurred more frequently in the previous context than W_j in W_j *is weak*.

The List Length Effect: In addition to the effect of practice, Anderson and Schooler (1991) studied the effect of the intervening interval on recurrence probability (see figure 2.8). Their results show a negatively accelerating curve which was best described as a power function where the exponent ranged between 0.73 and 0.83 depending on the type of material examined. Hence:

$$P(W_i \mid W_i \; in \; long \; interval) < P(W_j \mid W_j \; in \; short \; interval) \tag{5.2}$$

However, the affects of the interval between occurrences of an item may be a consequence of both the number of items in the context and the degree to which context has changed (c.f. Gillund & Shiffrin, 1984, analysis of context shifts and test delays in the experimental situation). The statistics collected by Anderson and Schooler (1991) were amalgamated over time by context. Three general contexts, namely, the *New York Times*, a database of children's verbal interactions, and mail messages received by John Anderson were analysed separately. While there may still have been some contextual drift within each broad context, the methodology would have minimised it. Further, in each of the analyses interval information was condensed using 100 item windows. Again this would work to minimise the amount of contextual drift. Finally, when Anderson and Schooler (1991) held the number of occurrences in the interval constant while they observed the affect of lag (in the spacing study), they found that the retention graphs flattened out significantly. Hence, it is likely that the decrement in recurrence probability observed was primarily a consequence of the number of additional items which are added to the context as the interval increases.

In addition, increasing the number of study items in recognition experiments decreases the average performance on the items, yet the affect of study list position is usually small (Gillund & Shiffrin, 1984). In particular, in the experiment which will be modelled (Ratcliff et al., 1990, experiment six) the only condition in which

a recency effect was found was the mixed weak, and in this case the effect may have been a consequence of the fact that at the end of the study list the weak items were surrounded mainly by familiar strong items (Ratcliff et al., 1990).

Hence, it is assumed that the number of unique items in a context rather than the contextual drift makes the primary contribution to inequality 5.2. That is:

$$P(W_i \mid W_i \textit{ in long context}) < P(W_j \mid W_j \textit{ in short context}) \tag{5.3}$$

where a long context contains many unique items and a short context contains few.

The List Strength Effect: The final effect which was investigated in the environment chapter (chapter two) was that of the list strength (i.e., the affect of repetition of other items on the recurrence probability of a given item). Unlike the item strength effect (which was strong) any list strength effect was not detected. If there is any affect of list strength it must be comparatively weak, and hence for the purposes of this study it will be assumed that it is non-existent.

Hence,

$$P(W_i \mid W_{k\neq i} \textit{ is weak}) \approx P(W_j \mid W_{k\neq i} \textit{ is strong}) \tag{5.4}$$

To summarise, there are three primary constraints to be considered: if an item is repeated in a context it is more likely to recur (i.e. inequality 5.1); if an item occurs in a long context which contains many unique items, it will be less likely to recur than an item from a shorter context (i.e. inequality 5.3); and, the probability of recurrence of an item is not affected by the amount of repetition of other items in a context (i.e. equality 5.4). The next section employs these constraints to construct a training set for presentation to the HRN.

5.3 Simulation: List strength and the HRN

The conditions for the list strength simulations most closely resembled the recognition conditions of Ratcliff et al.'s (1990) experiment six. In addition to the pure weak, mixed and pure strong lists which they had used in the previous list strength experiments, a length condition was added. Using this design it was shown that

the strength, length and null list strength effects could be obtained under the same conditions. The current simulations attempt to model these findings by constructing a training set which embodies the statistics outlined above, and training the HRN on this training set. There are three primary hypotheses:

Item Strength Effect: Items[1] which are presented multiple times should be recognised more accurately than items presented once.

Length Effect: Items from lists containing many unique items should be recognised less accurately than items from lists with few unique items.

Null List Strength Effect: The strength of list items other than the test item should not affect performance on the test item.

5.3.1 Method

Network Architecture and Parameters

Figure 5.1 shows the architecture which was used. All items were encoded locally at input and output. There were seventeen input units, one for each of the words and an extra "Pause" signal which indicated that there was no input at this timestep. As in the word frequency simulations the non zero value of each vector was set to 0.7 rather than 1 so that the targets were reachable (Touretzky & Pomerleau, 1991). Twenty two hidden units (and hence twenty two context units) and twenty output units were used. Three of the output units corresponded to "Blank" (i.e. do not output anything at this stage), "Yes" and "No" (i.e. was the test item in the study list or not?). The other seventeen outputs were the current inputs. As in the previous chapter, the input was autoassociated to ensure separation of the hidden unit patterns corresponding to different input patterns.

The network was trained with a learning rate of 0.05, a memory decay rate of 0.2 and a weight decay of 0.000002. Initial weights were selected from a uniform distribution between -1 and 1 and then updated after each pattern presentation.

[1] In the majority of the experimental work on the list strength effect in recognition the items used have been words. However, there is evidence that the same results obtain if the items are sentences (Murnane & Shiffrin, 1991b).

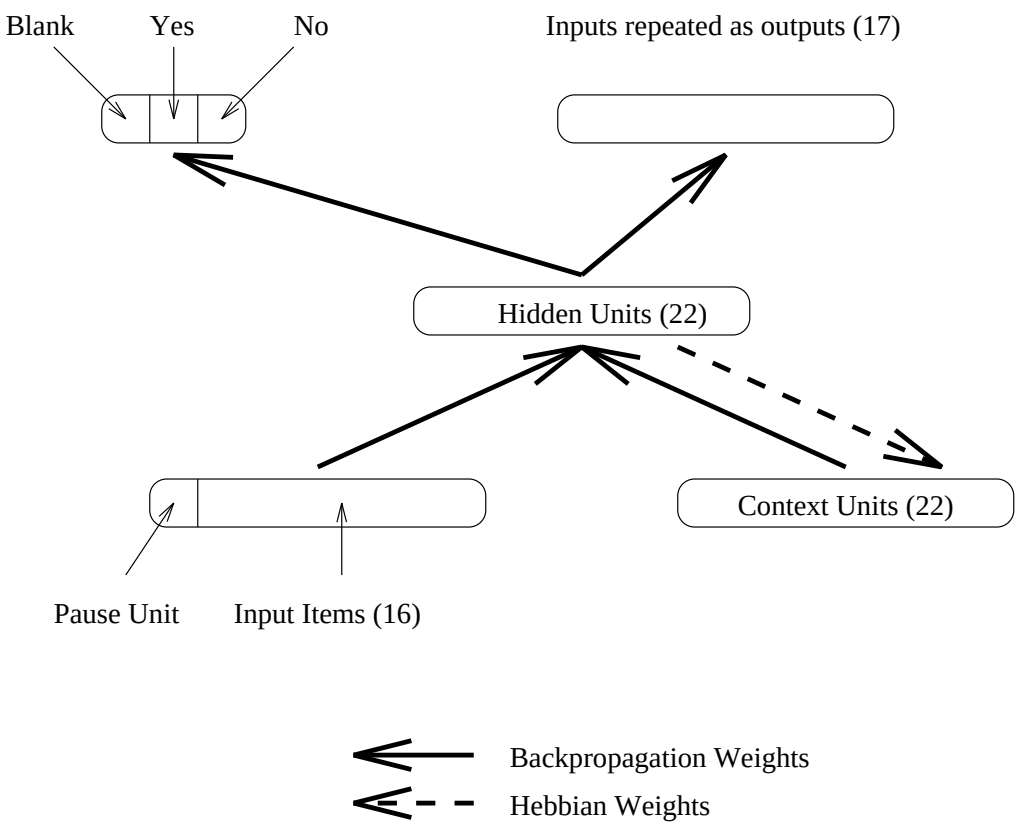

Figure 5.1: The HRN architecture for the list strength simulations. The architecture is a Hebbian Recurrent Network (HRN) as outlined in the HRN chapter (chapter three). The input vectors consisted of seventeen components. Sixteen of the input units represented words while one was a "pause" signal which indicated that no input was being presented at that time. The "pause" input was presented when the decision whether the test item was in the study list was being made. Of the outputs 17 were the inputs repeated and the other three coded for "blank", "yes" and "no". Either "Yes" or"No" was expected when the recognition decision was made and the "Blank" pattern was expected at all other times.

List Type	Condition	Study List	Test Item
Pure Weak	Pure Weak	G B	G
Mixed	Mixed Weak	C F C	F
	Mixed Strong	G G E	G
Pure Strong	Pure Strong	D J D J	J
Length	Length	H A I D B G L E	I

Table 5.1: Example target lists from the list strength training set.

The simulations were run for 1500 epochs by which time the performance on the simulations had reached asymptote.

Training Set Construction

As in chapter three, the episodic recognition task was simplified to presenting a list of items followed by a single test item. A "pause" pattern was then input and it was at this stage that the network was expected to respond (i.e. "Yes" or "No" depending on whether the item was in the list or not).

The first step in constructing a sequence was to decide on the study list type. There were four list types (i.e. pure weak, mixed, pure strong and length) each of which was chosen with the probability as indicated in table E.1. Pure weak lists consisted of two items presented once each. Mixed lists consisted of two items. One was presented once and the other twice. Pure strong lists also contained two items both of which were presented twice. The length condition contained eight items each presented once. Items were chosen with equal probability and the order of the study items was randomised. The test item was chosen with probabilities which depended upon condition (i.e. pure weak, mixed weak, mixed strong, pure strong and length). Table E.3 presents these probabilities and table 5.1 provides examples of each of the conditions.

The statistics of the training set are justified in the appendices. Appendix E.1 outlines how the environmental constraints (i.e. inequalities 5.1, 5.3 and 5.4.) can be re-expressed in terms of the training set constructs. Appendix E.2 presents the training set statistics while appendices E.3, E.4, and E.5 demonstrate that the training set statistics satisfy the item strength, list length and null list strength constraints respectively.

The only opportunity for the network to adjust to the environmental statistics is at test. Hence, it is important to ensure that there is a reasonable probability that each of the words occur as test items in each of the conditions in the training set. Because there are more conditions in the current simulations than in the word frequency simulations and since the probability of a given word occurring as a test item in a given condition is lower, the size of the training set was increased from 500 to 1000 sequences.

Restating the hypotheses in terms of the simulation conditions, it is expected that items in the strong conditions will be recognised more accurately than items in the weak conditions (item strength effect); items in the weak conditions will be recognised more accurately than items in the length condition (list length effect); and there will be no interaction between the mixed/pure and strength factors (null list strength effect). In addition to considering the interaction of the pure/mixed and strength factors, the analysis of the list strength effect has included the calculation of a ratio of ratios measure given by the following formula (Shiffrin et al., 1990; Ratcliff et al., 1990):

$$R_r = \frac{d'(mixed\,strong)/d'(mixed\,weak)}{d'(pure\,strong)/d'(pure\,weak)} \tag{5.5}$$

If weak items are not impaired by strengthening other items in the list (and conversely, if strong items do not improve by weakening other items in the list), the ratio of ratios measure will be near one. Ratio of ratios near or somewhat below one have been found in several studies (Ratcliff et al., 1990) and the simulations are expected to produce ratios near one.

Procedure

Twenty simulations were conducted. Each simulation was taken to correspond to a different subject. Initial weights, and training and test sets (containing 1000 sequences each) were generated for each simulation.

5.3.2 Results

For each network the *d'* values for each of the conditions were calculated from their hit and false alarm rates as indicated in chapter three (i.e. $d' = z_{hit\,rate} -$

$z_{false\ alarm\ rate}$). These *d'* values were then averaged across the simulations. Figure 5.2 shows the learning curves (i.e. *d'* as a function of the training epoch) for each of the conditions. A two way analysis of variance was conducted. Strong items were found to be better recognised than weak items, F(1, 19) = 238.83, p < 0.001, as illustrated by strong-to-weak ratios of 1.88 and 1.92 for pure and mixed lists respectively. There was no significant difference between pure and mixed lists, F(1,19) = 0.46, p = 0.50. Further, there was no significant list strength effect as assessed by the interaction term, F(1,19) = 0.36, p = 0.55. Figure 5.3 shows the list strength ratio as a function of training epoch. Before training, the ratio of ratios is well above one. But, for reasons outlined in the HRN chapter (chapter three), *d'* values which are calculated before training are not meaningful since there are three possible outputs. At 100 epochs the list strength ratio is below one since the performance on pure weak items lags behind that on the mixed weak items. As training progresses it rises to approximately one where it remains until 1500 epochs where it is 1.02.

Analyses of variance comparing the length condition against the two weak conditions were conducted. In both cases the performance in the length condition was significantly poorer than in the weak conditions (pure weak, F(1,19) = 8.32, p < 0.01; mixed weak, F(1,19) = 17.14, p < 0.001).

5.4 Discussion and Conclusions

As outlined in the results section the item strength and list length results were reproduced by the model while no list strength effect was observed. Hence, the three hypotheses were upheld in accordance with human experimental data (Ratcliff et al., 1990).

The performance of the model shows the same pattern as the probabilities of occurrence of each of the conditions in the training set (as outlined in appendix E). The item strength effect is generated by the fact that:

$$P(W_i \mid W_i\ is\ strong) > P(W_j \mid W_j\ is\ weak) \tag{5.6}$$

where W_i is a word within the training set. The length effect occurs because:

$$P(W_i \mid W_i\ in\ long\ context) < P(W_j \mid W_j\ in\ short\ context) \tag{5.7}$$

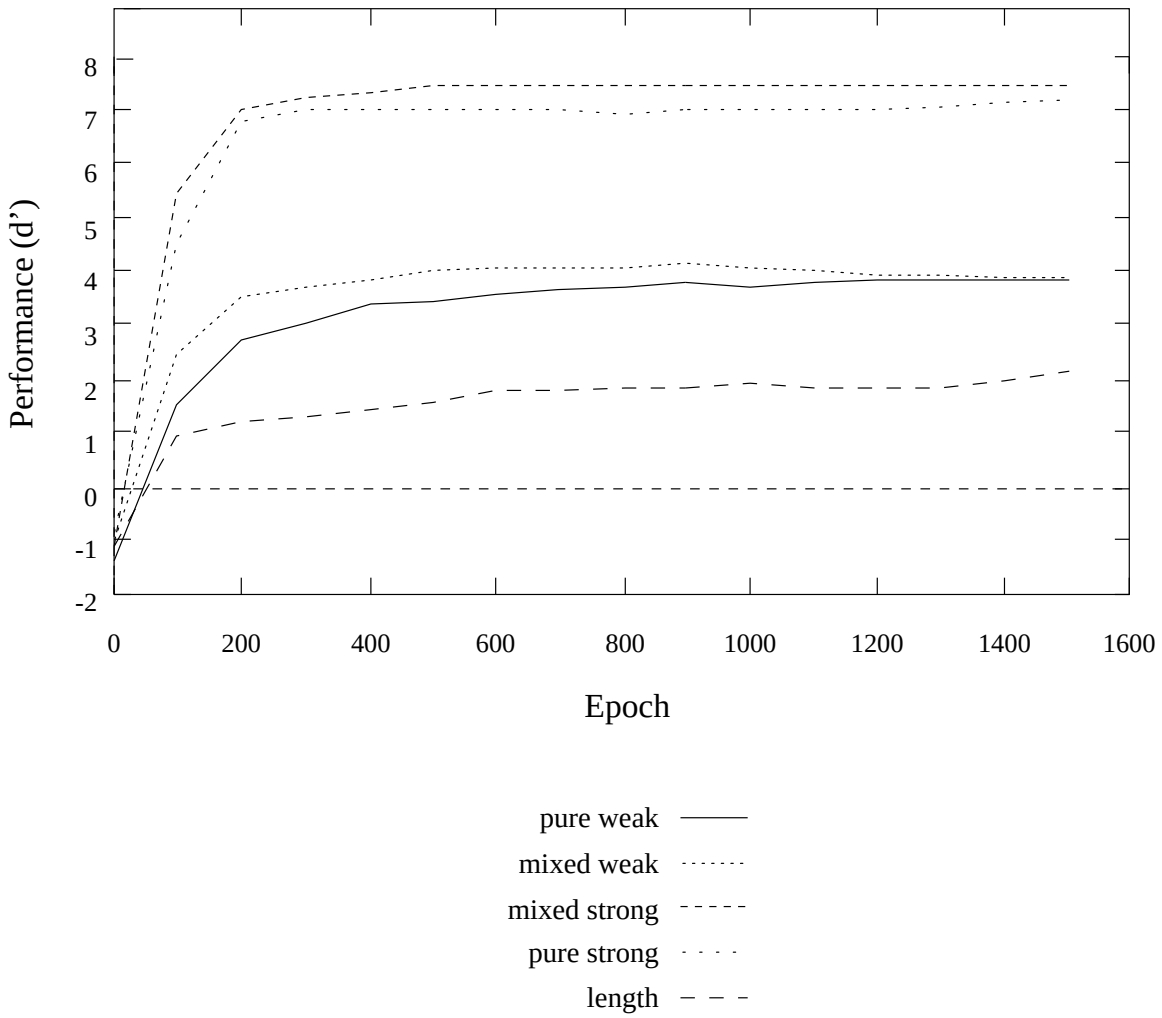

Figure 5.2: Training epoch verses performance (D') of the HRN for each of the list strength conditions. (i.e., pure weak, mixed weak, mixed strong, pure strong, length). The strong conditions show a better performance than the weak conditions which show better performance than the length condition. The performance on strong items is similar regardless of whether they came from mixed or pure lists. Also the performance on weak items is similar regardless of the list type from which they came. Hence the list strength ratio is near one.

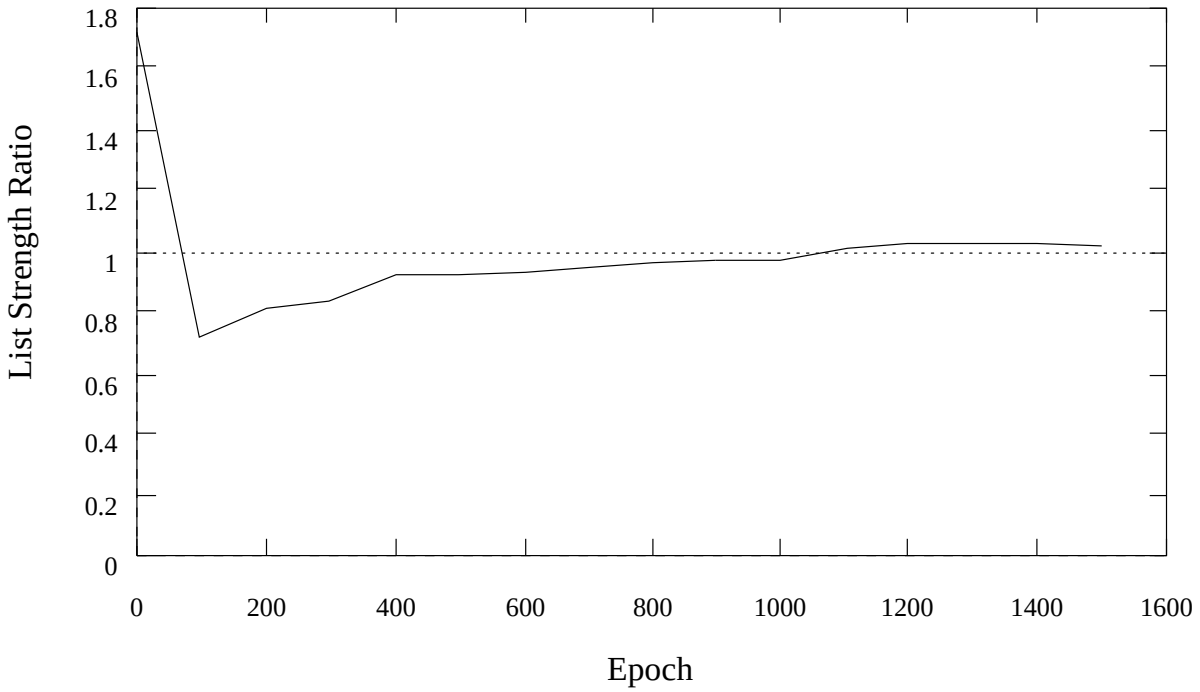

Figure 5.3: Training epoch verses the list strength ratio of the HRN. The list strength ratio rises gradually to near one indicating that there is no list strength effect.

and the null list strength effect obtains since:

$$P(W_i \mid W_{k \neq i} \; is \; weak) \approx P(W_j \mid W_{k \neq i} \; is \; strong) \tag{5.8}$$

In each case, these constraints reflect the findings of the environmental analysis and hence, the suggestion is that the human memory system demonstrates the pattern of performance results that it does because it is optimised to the environmental frequencies of recurrence.

While the matrix model component of the HRN could account for both the item strength and length effects, it incorrectly predicts a strong list strength effect. In the matrix model, however, the decision criterion is a thresholded dot product. In contrast, the decision mechanism of the HRN is determined by the backpropagation weights which are optimised to the environment. As was determined in the environment chapter (chapter two), however, the strength of other items within a context makes no significant difference to the probability of an item recurring in the environment. Hence, training with a data set which incorporates these constraints produces decision boundaries which are positioned to minimise the affect of other items[2].

[2]Analysing the placement of decision boundaries in the trained HRN is difficult due to their

In discussion of the word frequency effect, the independent variable, frequency, is an environmental variable. It seems quite reasonable and perhaps even necessary[3] to invoke an environmental explanation of some description. For instance, Humphreys et al. (1989) allude to the possibility that part of the information that is available to the memory system about a word is its frequency. Presumably this information is assembled by the mechanism as it is exposed to the environment.

By contrast, in accounting for the pattern of results arising from strength and length manipulations the impact of the environment is not as obvious. None of the independent variables are necessarily environmental, and the explanations of the null list strength effect that have arisen (Shiffrin et al., 1990; Chappell & Humphreys, in press, 1993; Chappell, 1993; Murdock & Kahana, 1992; Heathcote, 1993) have concentrated on mechanism (see chapter two for a review). However, what this study demonstrates is that even a phenomenon such as the null list strength effect can be explained by a simple learning mechanism exposed to an observable environment.

high dimensionality. Unlike points which can be analysed using techniques such as principal components analysis or canonical discriminants analysis, hyperplanes are more difficult to reduce to low dimensions without loosing important information.

[3]Provided one believes that the effect occurs because of the frequency of the word and not as a consequence of some accidently correlated variable such as word length or concreteness.

Chapter 6

Conclusion and Further Work

6.1 Summary of the Contribution

This thesis describes an exploration of the viability of the interactivist approach. I argue that performance on memory tasks can be modelled as an interaction between the *mechanism* of memory and the *environment* of memory. The aim was not to provide a detailed psychological model of the processes of human memory, nor to determine a definitive methodology for empirically studying the environment. The point was to identify the issues involved in implementing interactivism. The thesis provides an existence proof that such a mechanism is possible and that useful statistics can be obtained from the environment.

In chapter one, I proposed that producing an interactive account of a phenomenon involves answering three questions:

1. Can the environment relevant to the phenomenon be adequately characterised?
2. Can a model that integrates learning into memory be constructed?
3. When the learning mechanism is considered in the context of the environmental statistics does it produce the behaviour that is found in the experimental setting?

The first of these questions was addressed in chapter two where I demonstrated that it is possible to characterise the environment of some phenomena by providing analyses of the relevant environments of the word frequency and null list strength

effects in recognition. In the case of word frequency, it was found that although high frequency words occur more often than low frequency words, low frequency words were more likely to recur within a context. In the null list strength case, it was found that items that are repeated within a context have higher recurrence rates and that the degree of repetition of items other than the target item had no significant affect on its recurrence rate.

The second question requires the construction of a model that integrates learning into memory. Chapter three presented such a model, called the Hebbian Recurrent Network (HRN), which incorporates the matrix model (Pike, 1984; Humphreys et al., 1989) into the Simple Recurrent Network (SRN, Elman, 1989, 1990). The result is an architecture that has the desirable memory characteristics of the matrix model, such as low interference and massive generalisation, but which is able to *learn* appropriate encodings for items, decision criteria and the control functions of memory, which have traditionally been chosen *a priori* in the mathematical memory literature. Unlike other recurrent architectures, the HRN does not need to be taught to retain information over long durations. Its architecture ensures that information is stored. The task is to learn the encoding and retrieval operations. Simulations demonstrate that the HRN is well suited to a recognition task inspired by typical memory paradigms. When compared with the SRN, the HRN is able to learn longer lists, generalises from smaller training sets and is not degraded significantly by increasing the vocabulary size.

The third question involves the integration of the environmental statistics and the learning mechanism to provide an account that addresses the experimental data. Two case studies of this synthesis were outlined in chapters four and five. In chapter four, the HRN was found to be sensitive to the higher recurrence rate of low frequency words. When the HRN was exposed to a training set in which low frequency words recurred with greater probability than high frequency words, it showed a low frequency word advantage, as is seen in human subjects (Gorman, 1961). In a similar way, chapter five demonstrated that, given the appropriately constrained training set, the HRN shows the item strength, list length and null list strength effects. Hence, even in phenomena in which none of the independent variables are directly environmental, an interactive account that implicates both the mechanism

and the environment can be constructed, thus providing some confidence in the generality of the approach.

6.2 Evaluation and Further Work

The introduction of learning into models of memory is important for the study of memory and its role within the developing cognitive architecture. In the introduction (chapter one), I outlined some important issues that can be addressed by models of memory that learn. In the following sections, these issues are re-examined in light of the interactive framework, the environmental analysis of the word frequency and list strength effects, and the Hebbian Recurrent Network. The synthesis of the environmental and architectural components into a coherent framework has implications for further work, which are also discussed in the following sections.

6.2.1 Learning and Development

One striking feature of the literature on human memory is the division that exists between the work on memory development and that on adult memory performance. The literature on memory development has been largely descriptive and computational models are in their infancy (Kail, 1984). Adopting Bickhard's position (see appendix A), one might argue that such a division is a direct consequence of the encoding style representations, which adult memory theorists have implicitly adopted. By contrast, the interactivist perspective (Bickhard, 1991; Bickhard & Campbell, 1993; Bickhard & Terveen, 1993) was developed specifically to address developmental issues and, hence, is suited to bridging the gap between adult and developmental memory models.

Producing a computational model of memory development requires introducing a learning component into memory. The HRN architecture is an explicit attempt to achieve such an integration. In particular, the HRN highlights the necessity of including memory specific predispositions to store information (c.f. Karmiloff-Smith, 1992a, 1992b, 1992c). In the HRN, such a predisposition is achieved by adding Hebbian weights to the SRN architecture. These Hebbian weights do not have to be trained to store information, rather, they automatically retain whatever appears on

the hidden units. The linear time complexity of learning in the HRN (chapter three) suggests that performance limitations apparent in the SRN as vocabulary and list length are increased have been avoided.

Given the previous discussion, further work involving the application of the HRN to developmental phenomena may be warranted. While recognition memory, present in the first few months of life (Fagan, 1973), seems to be a very basic ability, the skills (i.e. control processes) involved in applying strategies take much longer to develop (Kail, 1984). It may, however, be possible to train the HRN to perform such strategies (e.g. rehearsal, categorisation) and, therefore, provide computational models of developmental memory phenomena.

6.2.2 Learning-to-Learn

The HRN differs from current mathematical models of human memory in that it makes learning-to-learn its primary focus. In the HRN, the representations, decision criteria and control processes are all learned. While the mathematical models (reviewed in chapter one) assume that the representations of items form a similarity landscape, the HRN provides a mechanism by which such a landscape might be learned. Whereas current models provide decision criteria that are fixed in form, the HRN presents a mechanism by which complex decision criteria are developed. Whereas current models specify the control processes, the HRN provides a mechanism by which control processes are acquired in response to the demands of the environment. The HRN emphasizes the mechanisms by which these memory components are formed rather than the components themselves.

One advantage of approaching memory from a learning perspective is that it focusses attention on several issues that have been difficult to address in the information processing paradigm. The origin of the similarity landscape of representations is one example that has already been mentioned. Perhaps the most prevalent area, though, is control processes. Processes such as rehearsal or list categorisation have been assumed to affect the memory trace, but the construction of a computational model of these processes has proven to be a formidable task. At least part of the difficulty has been the "virtually unlimited" (Atkinson & Shiffrin, 1968, p. 106) variety of control processes that subjects apply. Using the information processing

List 1	
Table	Damp
Beach	Lame
Sing	Dark
List 2	
Beach	Weak
Sing	Black
Table	Moist

Table 6.1: Example of lists in the ABAB' paradigm. The second items in the first list are replaced with synonyms to form the second list.

approach, one would have to outline the algorithms of each of these processes on a case by case basis. In contrast, the HRN supposes that these processes are learned. In an interactive account, the variety of subject strategies could be ascribed to the variety of environmental demands while maintaining a relatively simple mechanism. Investigating this hypothesis, using the HRN as the interactive mechanism, may provide a fruitful basis for additional work.

Furthermore, if the interactivist account has merit and phenomena such as the word frequency effect and the null list strength effect are a consequence of adaption to the environment in the way that has been suggested in this thesis, it follows, that if the environment is changed the behaviour must also be altered. The work presented here demonstrates how an interactive account of memory phenomena can be constructed, but an experiment in which a phenomenon such as the list strength or word frequency effect was altered by training would provide a strong case for a more dynamic perspective on memory.

Phenomena that change qualitatively with training have been reported in the literature. In the ABAB' paradigm, a list of pairs is presented followed by a second list in which the first item of each pair in list one is re-paired with a synonym of the second item (see table 6.2.2). Naive subjects perform more poorly on the synonym list than a control list in which none of the items were repeated (i.e. the ABCD paradigm). That is, the synonyms interfere with the subjects' ability to recall the correct item. As training progresses, the result changes and the synonyms facilitate recall relative to the control (Postman, 1964). In addition, the performance on ABAC and ABABr lists, which both show negative transfer initially, was found

to improve more quickly than that on the ABCD lists. One would assume that the subjects were not only learning the appropriate response set, but were also learning to form more efficient associations (Postman, 1969). If one were to model these paradigms in the HRN, the context-to-hidden and hidden-to-output weights would be responsible for response learning, while the input-to-hidden and context-to-hidden weights would account for the learning that seems to occur in the association formation process (see figure 3.2).

Suggested Experiment: The non-specific transfer studies described above demonstrate that the performance on memory tasks changes as a consequence of training and that this change can be qualitative as well as quantitative. Exposing subjects to an environment in which the statistics of occurrence are systematically different from those in their normal environment may lead to a change in their behaviour. For instance, in the case of the word frequency phenomenon, subjects would be exposed to an experimental environment in which high frequency words were more likely to recur than low frequency words. The prediction would be that they would show a reduced advantage for low frequency words or perhaps a high frequency advantage. In the list strength case, subjects would be exposed to an environment that has been structured so that the probability of occurrence depended on the degree of repetition of other items. In this way, it may be possible to induce a list strength effect in recognition. If behaviour was altered, there would be a strong argument for expanding the role of the environment in models of human memory and for adopting a mechanism such as the HRN in which both the association forming mechanism and the response mechanism involve learning. Tests of these hypotheses, however, are beyond the scope of this thesis.

6.2.3 Learning and Meaning

Interactive accounts of cognition circumvent the problem of establishing the content of representations (Bickhard, 1991). In many current models, including TODAM (Murdock, 1982), CHARM (Eich, 1982), SAM (Gillund & Shiffrin, 1984), Minerva II (Hintzman, 1984), the matrix model (Pike, 1984; Humphreys et al., 1989) and rational analysis (Anderson & Milson, 1989), representations are assumed to *cor-*

respond to objects and states in the environment. Complex representations (such as associations) are composed using a fixed operation such as concatenation, vector outer product or convolution. Bickhard (1991) argues that such representational schemes could not have developed either within an individual or by the processes of evolution (see Appendix A). While it is difficult to determine with certainty whether the preconditions of Bickhard's argument apply to these models, with the exception of Hintzman (1986), memory theorists have not addressed how the representations they use attain their content. In contrast, the content of interactive representations, such as those employed in the HRN, can be traced to the dynamic differentiation of environments. In this way, the encoding problem outlined by (Bickhard, 1991) can be avoided. Furthermore, it has been demonstrated that some forms of meaning can be feasibly extracted from the environment using the SRN (Elman, 1989, 1990, 1991), on which the HRN is based. Elman (1990) used a small artificial grammar to generate sentences. These sentences were then filtered to satisfy certain semantic constraints. For instance, only animate nouns appeared as the subject of "eat" and only edible substances appeared as the object. The network was trained on a prediction task with this corpus and the resulting hidden unit representations encoded both syntactic roles (such as verb or noun) and semantic features (such as animal or edible).

A possibility for further work is to train the HRN on a corpus like that chosen by Elman (1989, 1990, 1991) or reduced versions of text databases such as those used in the environmental analyses in chapter two. The performance of the HRN could then be compared against the experimentally observed affects of similarity (see Osgood (1949), Martin (1965) and Eich (1982) for reviews of the similarity literature and Eich (1982) for an attempt to model the affects in CHARM).

6.2.4 Learning and Cross Paradigm Generality

An advantage of employing a connectionist learning mechanism is the degree of cross paradigm generality that it provides. While most memory models are designed specifically for laboratory paradigms, such as recognition or cued recall, the HRN employs no task specific predispositions. Instead, it acquires the ability to perform a task as a consequence of exposure to an environment in which the task

is embedded. In principle, the HRN architecture could perform categorisation, inferential reasoning or problem solving. In this thesis, however, the simulations have been restricted to recognition memory, and in the next section the parsimony with which the HRN accomplished the word frequency and list strength findings is examined. While there are many sets of results to which the HRN could be applied, the section entitled "Extension to other paradigms and results" outlines two of the most immediate possibilities, namely, cued recall and reaction time paradigms. Extensions such as these belong to what might be called a *task centered* approach in which the emphasis is on modelling, in detail, a well established set of laboratory tasks. Some dangers of such an approach are outlined in the section entitled "The problem with focussing on tasks". The following section (i.e., Principle focussed modelling: Encoding specificity) presents an alternative called the *principle centered* approach in which the emphasis lies on modelling principles that seem to underlie large sets of cognitive tasks both from the laboratory and in naturalistic settings.

Parsimony

In chapters four and five, the HRN was applied to the word frequency and null list strength effects respectively, allowing the parsimony offered by an interactive approach to be evaluated. In modelling the word frequency effect, the free parameters were the learning rate, the memory decay rate, the weight decay rate and the number of hidden units[1]. In the modelling of the item strength, length and null list strength effects, no extra parameters were added to the mechanism.

Further, the learning rate and memory decay rate remained unaltered. The only model parameters that were changed were the number of hidden units and the weight decay. These are the coarse and fine controls of a single parameter, namely the number of effective parameters. Hence, the ability of the HRN mechanism to capture phenomena without expanding the number of parameters seems favourable, although, it is too early to make any firm conclusions. When compared against the matrix model, on which the HRN is built, the interactive explanations of the word

[1]In error correcting models of cognition there are two levels of optimisation taking place. The backpropagation algorithm is optimising the weights of the network to the training set and the experimenter is optimising the parameters of the network to the experimental data. In evaluating the parsimony of the model it is the latter parameters that are of interest.

frequency and list strength effects gain much by appealing to the explanatory power of the environment statistics. While the matrix model can account for both the affect of item strength and list length, the advantage for low frequency words and the null list strength effect were generated by the environment. This explanatory power underlies the promise of the interactive approach.

Extension to Other Paradigms and Results

Recall: As mentioned in the learning-to-learn section above, cued recall would be a useful place to begin extending the HRN to other paradigms. Unlike serial and free recall, which require significant control, cued recall could be handled by presenting the network with both items of the study pairs. The input-to-hidden weights would learn to form associations of these items and the Hebbian weights would store these associations. Subsequently, the system could be cued with either of the items. Such an extension should be possible without any alteration of the model.

Both serial recall (Phillips, 1991) and free recall (Nolfi et al., 1990) have been modelled using the SRN. In each case, however, both the vocabulary size and the list length have been severely restricted. As was noted above, adding Hebbian weights can greatly improve both the storage capacity and the ability to process a large vocabulary indicating that the HRN may perform well on these tasks without any alteration of the architecture.

Reaction time: The HRN model does not produce a measure of reaction time. Reaction time is an important ingredient in model identification (Anderson, 1990) as demonstrated by the distributions for both recognition and cued recall that have been collected recently (see Nobel & Shiffrin, 1992). There are several possible extensions to the HRN that would provide a measure of reaction time. One possibility is to use a time averaged net input to the activation function (Cohen, Dunbar, & McClelland, 1990; Reilly, 1993). That is,

$$net_j(t+\Delta t) = \tau \sum_i a_i(t+\Delta t) w_{ji}(t+\Delta t) + (1-\tau) net_j(t) \tag{6.1}$$

where $net_j(t)$ is the net input to unit j at time t, τ is a decay parameter, $a_i(t)$ is the activation of unit i at time t, and $w_{ij}(t)$ is the weight between unit i and unit j at time t.

Another possibility is to time average the activation itself (Cottrell, Nguyen, & Tsung, 1993), that is:

$$a_j(t+\Delta t) = \tau f(\sum_i a_i(t+\Delta t)w_{ji}(t+\Delta t)) + (1-\tau)a_j(t) \tag{6.2}$$

where f is the asymmetric sigmoid.

To incorporate a reaction time measure and maintain the interactive principle is more challenging, however, because one would require a model that not only attempted to give the optimal answer, but actively adapted to give it at the optimal time relative to the environment. To incorporate the interactive perspective would require not only an extension to the model but also an extension of the environmental analysis to record the amount of time that the memory system has to respond.

In this section, a number of tasks have been suggested that could be modelled with minor modifications to the HRN. In addition to expanding the number of results for which the model can account, however, it is important to consider how the model illustrates general principles of cognition. The next section outlines the difficulties with restricting oneself to an approach that focusses on tasks, while the following section outlines an approach centered on principles of cognition as another way in which work on the HRN might be extended.

The Problem with Focussing on Tasks

It has been argued that many connectionists have concentrated on modelling tasks rather than the principles underlying cognition or cognitive development as a whole (Karmiloff-Smith, 1992c; Smolensky, Legendre, & Miyata, 1992). Such an approach can prevent insights that arise from the consideration of cross domain influences. As an example, Karmiloff-Smith (1992c) describes experiments in which French speaking children are presented with article/noun pairs in which the noun is meaningless yet phonotactically legal. Subsequently, they were observed to see if they ascribe the noun with a masculine or feminine gender. Interestingly, children between the ages of three and seven ignored both the natural gender of the object with which they were presented and the gender of a preceding article (a foolproof indication of the gender), and focussed on the word ending despite the fact that word endings do not always indicate gender. However, when Karmiloff-Smith trained a three layer perceptron with backpropagation on this task, the network performed faultlessly. It

concentrated primarily on the reliable article cue and never made the mistakes of the three to seven year olds. She suggests two ways in which the model may have differed from the children: the first is the possibility that people are genetically predisposed to concentrate on word endings; the second proposed that the definition of the environment was too specific to the gender task that she was modelling. The task before children is to acquire an entire language including many other rules such as determining the past tense of verbs in which word endings are crucial. She concludes;

> "most expert Connectionists have also not really modelled development, they have modelled tasks ... However, if we are to model the content of children's learning in specific domains, then models should ultimately reflect the complexities of the child's interaction with the world and the influences of the timing and patterning of cross-domain learning."

(Karmiloff-Smith, 1992c, p 263)

In the discussion of the list strength effect, the importance of modelling the pattern of phenomena (i.e. the strength, length and null list strength effects) was noted. However, all of the simulations in this thesis have concentrated on the single item recognition task. The next section outlines an alternative approach and discusses a useful starting point in such an endeavour, namely, the *encoding specificity principle*.

Principle focussed modelling: Encoding specificity

An alternative to concentrating on additional tasks is to attempt to illuminate principles that underlie many cognitive tasks (Smolensky et al., 1992). One such principle is the importance of the similarity of encoding and retrieval processes in facilitating retrieval (*encoding specificity*, Tulving & Thomson, 1973). Information is constantly being processed, stored and retrieved by the human cognitive system. Successful retrieval depends upon the formation of an appropriate cue with which to probe memory. What suffices as an appropriate cue, in turn, depends upon the encoding process that was used to form the stored representation. There is a mutual dependence between the encoding and retrieval operations that underlie many cognitive phenomena, including comprehension, inference, problem solving and analogical reasoning.

Consider, for instance, the comprehension of a simple paragraph such as:

John went to the zoo. He saw an elephant.

To test the comprehension, the subject might be queried with questions such as:

Where did John go?
What did John see?
Who saw an elephant?
How old is John?

The subject encodes a set of "propositions" while reading the paragraph and stores these in memory (Kintsch & van Dijk, 1978). When the time comes to answer the questions, they must form appropriate cues to retrieve the information and then output the required component. The question is how the encoding and retrieval operations are coordinated despite the fact that there may be a large gap between the encoding event and the retrieval event.

Cognitive models that suppose that memory processes are primarily determined by a fixed cognitive architecture ensure that these processes interact appropriately by design. For instance, distributed memory models of single item recognition, such as TODAM (Murdock, 1982) or the matrix model (Pike, 1984; Humphreys et al., 1989), presume that the pattern or vector that is formed and stored as a consequence of the encoding processes is similar to the pattern that is formed as a retrieval cue, despite the fact that these processes may occur through different modalities, in very different circumstances, and may be separated by a considerable timespan. It is this similarity that ensures that the retrieval operation (i.e. the dot product in this case) is appropriate.

This thesis, however, has been an exploration of the hypothesis that the representations, decision criteria and control processes of the cognitive system are learned rather than designed. The question of how encoding and retrieval processes are coordinated is an issue for further investigation. However, the gains in parsimony and cross paradigm generality that accrue by adopting a learning model and allowing the environment to generate much of the explanatory power suggest that such investigation is warranted.

6.2.5 Learning and the Environment

The major emphasis of this thesis is the role that it outlines for the environment in models of human memory. The analyses in chapters two, four and five demonstrate that the environment is structured in such a way that a memory system that learns can produce the word frequency and null list strength effects in recognition. While these studies do not show a causal link between the nature of the environment and the performance of the memory system, they establish that the appropriate preconditions do exist for explanations of these phenomena that presume that the environment is not simply the source of inputs and the destination of outputs, but plays a role in determining the nature of the memory system (Anderson, 1990).

Methodologically, the most important consequence of the interactive perspective is that the environment must be studied empirically. The word frequency and list strength examples demonstrate that the collection of environmental statistics in the context of an optimising memory system such as the HRN can provide insight into why memory performs as it does.

There is, however, still much to do in naturalising the training set. The sets on which the HRN was trained involved repetitions of the target task, episodic recognition. While it seems quite likely that recognition is a part of the reading process, subjects probably don't cycle through the words in a newspaper subvocalizing "old, new, new, old, new". Most of the subjects studied in experimental settings have been involved in formal schooling, which may include tasks similar to those encountered in the laboratory. It is unlikely, however, that these subjects are exposed to any formal experimental situation prior to having to perform the memory tasks required by the experimenter.

Another way in which the training sets depart from those seen in the typical subject's history, is that negative evidence is supplied. The sequences presented to the network included ones in which the test item did not occur and the network was require to output "No" in these cases. In contrast, the acquisition of language and memory skills occurs largely in the absence of negative evidence. Connectionist models of memory must include the capability to learn from positive examples alone[2].

[2]MacWhinney (1993) and Regier (1992) discuss ways in which connectionist networks might

The case studies in chapters four and five take a first step in ensuring that the training set reflects the environmental landscape in a qualitative fashion. However, constructing models that learn on the basis of training sets that more closely resemble the everyday experience of the subject will be necessary to provide an ecologically faithful picture of the memory system.

This thesis examines the consequences of introducing learning into models of human memory. There are two major implications for memory research. The first is a change from a methodology that has equated memory research with the cataloging of memory processes. Each of the major models includes a model of recognition, a model of recall and a model of familiarity judgement to name a few. In a learning model such a catalog is replaced by an interest in the mechanism by which these memory processes come into being. The second implication is the role of the environment that a learning model necessitates. The environment becomes an active participant in the structuring of the memory system, and must be examined empirically to account for human memory phenomena.

take advantage of implicit negative evidence to circumvent the absence of negative examples in the environment.

Bibliography

Allen, L. R., & Garton, R. F. (1968). The influence of word knowledge on the word frequency effect in recognition memory. *Psychonomic Science*, *10*, 401–402.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413–451.

Anderson, J. R. (1989). A rational analysis of human memory. In Roediger, L., & Craik, F. I. M. (Eds.), *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*, pp. 195–210. Lawrence Erlbaum, Hillsdale, NJ.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Laurence Erlbaum Associates, Hillsdale, NJ.

Anderson, J. R. (1991). Is human cognition adaptive. *Behavioral and Brain Sciences*, *14*, 471–517.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K. W., & Spence, J. T. (Eds.), *The Psychology of Learning and Motivation*, pp. 90–191. Academic Press.

Barnes, J. M., & Underwood, B. J. (1959). Fate of first-list associations in transfer theory. *Journal of Experimental Psychology*, *58*, 97–105.

Bates, E. A., & Elman, J. L. (1992). Connectionism and the study of change. Technical Report 9202, Center for Research in Language, University of California, San Diego, CA.

Belew, R. K. (1990). Evolution, learning, and culture: Computational metaphors for adaptive algorithms. *Complex Systems*, *4*, 11–49.

Bickhard, M. H. (1991). The import of Fodor's anti-constructivist argument. In Steffe, L. P. (Ed.), *Epistemological Foundations of Mathematical Experience*, chap. 2, pp. 14–25. Springer Verlag, New York, NY.

Bickhard, M. H., & Campbell, R. L. (1993). Topologies of learning and development. Unpublished manuscript.

Bickhard, M. H., & Terveen, L. (1993). Impasse and solution: Foundational issues in artificial intelligence and cognitive science. Unpublished manuscript.

Blough, D. S., & Millward, R. B. (1965). Learning: Operant conditioning and verbal learning. *Annual Review of Psychology*, *16*, 63–94.

Bower, G. H. (1972). Mental imagery and associative learning. In Gregg, L. W. (Ed.), *Cognition in Learning and Memory*, pp. 51–88. Wiley, New York.

Brannelly, S., Tehan, G., & Humphreys, M. S. (1989). Retrieval plus scanning: Does it occur?. *Memory and Cognition*, *17*(6), 712–722.

Brousse, O., & Smolensky, P. (1989). Virtual memories and massive generalization in connectionist combinatorial learning. In *Program of the Eleventh Annual Conference of the Cognitive Science Society*, pp. 380–387 Hillsdale, NJ. Lawrence Erlbaum Associates.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley, CA.

Campbell, D. T. (1974). Evolutionary epistemology. In Schilp, P. A. (Ed.), *The philosophy of Karl Popper*. Open Court, La Salle, IL.

Campbell, R. L., & Bickhard, M. H. (1991). If human cognition is adaptive, can human knowledge consist of encodings?. *Behavioral and Brain Sciences, 14* (3), 488–489.

Carpenter, G. A., & Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organizing neural network. *Computer, March*, 77–88.

Chalmers, D. J. (1990). Syntactic transformations on distributed representations. *Connection Science, 2*(1 & 2), 53–62.

Chappell, M. (1993). *Neural Network Models of Human Recognition and Cued Recall.* Ph.D. thesis, The University of Queensland.

Chappell, M., & Humphreys, M. S. (1993). The null list-strength effect can be modelled by a neural network. In Goss, S., Caelli, T., Oldfield, S., & Wales, R. (Eds.), *Proceedings of the Second Australian Cognitive Science Conference*, pp. 46–48.

Chappell, M., & Humphreys, M. S. (in press). Analysis of a neural network with application to human memory modelling. *Journal of Experimental Psychology.*

Charniak, E., & McDermott, D. (1985). *Introduction to Artificial Intelligence* (World Student Series edition). Addison Wesley, Reading, MA.

Chauvin, Y. (1989). A backpropagation algorithm with optimal use of hidden units. In Touretsky, D. S. (Ed.), *Advances in Neural Information Processing Systems*, pp. 519–526. Morgan Kaufmann.

Cohen, J. D., Dunbar, K., & McClelland, J. M. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97*, 332–361.

Collins, A., & Smith, E. E. (Eds.). (1988). *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence.* Morgan Kaufmann Publishers, San Mateo, CA.

Cottrell, G. W., Nguyen, M., & Tsung, F. S. (1993). Tau net: The way to do is to be. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* Hillsdale, NJ. Lawrence Erlbaum Associates.

Cottrell, G. W., & Tsung, F. S. (1993). Learning simple arithmetic procedures. *Connection Science*, *5*(1), 37–58.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports*, *7*, 337–344.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological errors. *Cognitive Science*, *17*, 149–195.

Dennis, S., Wiles, J., & Humphreys, M. S. (1992). What does the environment look like? Setting the scene for interactive models of human memory. Technical Report 249, Department of Computer Science, The University of Queensland, Australia.

Duncan, C. P. (1960). Descriptions of learning to learn in human subjects. *American Journal of Psychology*, *73*, 108–114.

Ebbinghaus, H. (1964). *Memory: A Contributon to Experimental Psychology*. Dover Publications, New York, NY. Translated by H A Ruger and C E Bussenius. Original written in 1885.

Eich, J. M. (1982). A Composite Holographic Associative Recall Model. *Psychological Review*, *89*(6), 627–661.

Elman, J. L. (1989). Representation and structure in connectionist models. Technical Report 8903, Center for Research in Language, University of California, San Diego, CA.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J. L. (1991). Incremental learning, the importance of starting small. Technical Report 9101, University of California, San Diego, CA.

Estes, W. K. (1991). Cognitive architectures from the standpoint of an experimental psychologist. *Annual Review of Psychology*, *42*, 1–28.

Fagan, J. F. (1973). Infants' delayed recognition memory and forgetting. *Journal of Experimental Child Psychology*, *16*, 424–450.

Feldman, J. A., Lakoff, G., Stolcke, A., & Hollbach-Weber, S. (1990). Miniature language acquisition: A touchstone for cognitive science. Technical Report TR-90-009, International Computer Science Institute.

Fodor, J. (1980). On the impossibility of acquiring more powerful structures. In Piatelli-Palmarini, M. (Ed.), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, pp. 142–162. Harvard University Press, Cambridge, MA.

Fodor, J. A. (1981). The present status of the innateness controversy. In *Representations*, chap. 10, pp. 257–316. The Harvester Press.

Fodor, J. A., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Francis, W. N., & Kucera, N. (1982). *Frequency analysis of English Usage: Lexicon and Grammar*. Hought Mifflin Company, Boston, MA.

French, R. (1991). Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. Technical Report 51, Center for Research on Concepts and Cognition, Indiana University, IN.

Gardner, M. K., Rothkopf, E. Z., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory and Cognition*, *15*(1), 24–28.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Gigerenzer, G. (1991). Does the environment have the same structure as Bayes theorem?. *Behavioral and Brain Sciences*, *14*, 495–496. Comment on Is human cognition adaptive? by J R Anderson in same volume.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*(1), 5–16.

Glanzer, M., Adams, J. K., & Iverson, G. (1991). Forgetting and the mirror effect in recognition memory: Concentering of underlying distributions. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *17*(1), 81–93.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 21–31.

Gorman, A. N. (1961). Recognition memory for names as a function of abstractness and frequency. *Journal of Experimental Psychology*, *61*, 23–29.

Greene, R. L. (1992). *Human Memory: Paradigms and Paradoxes*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Greeno, J. G., & Bjork (1973). Mathematical learning theory and the new "mental forestry". *Annual Review of Psychology*, *24*, 81–116.

Gregg, V. H. (1976). Word frequency, recognition and recall. In Brown, J. (Ed.), *Recall and recognition*. Wiley, London.

Hadley, R. F. (1993). Compositionality and systematicity in connectionist language learning. Technical Report CSS-IS TR 93-01, School of Computing Science and Cognitive Science Program, Simon Fraser University, Canada.

Hall, J. F. (1954). Learning as a function of word frequency. *American Journal of Psychology*, *67*, 138–140.

Harnad, S. (1990). The symbol grounding problem. In Forrest, S. (Ed.), *Emergent Computation*, pp. 335–346. North-Holland.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*, 356–388.

Heathcote, A. (1993). An ART model of human recognition memory. In Leong, P., & Jabri, M. (Eds.), *Proceedings of the Fourth Australian Conference on Neural Networks*, pp. 212–215.

Hetherington, P., & Seidenberg, M. S. (1989). Is there catastrophic interference in connectionist networks?. In *Program of the Eleventh Annual Conference of the Cognitive Science Society*, pp. 26–33 Hillsdale, NJ. Lawrence Erlbaum Associates.

Hinton, G. (1993). Tutorial on neural networks. Conducted at University of Sydney.

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, and Computers*, *16*(2), 96–101.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.

Hintzman, D. L. (1991). Why are formal models useful in psychology?. In Hockley, W. E., & Lewandowsky, S. (Eds.), *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*, pp. 39–56. Lawrence Erlbaum Associates, Hillsdale, NJ.

Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multi-trace hypothesis. *Journal of Experimental Psychology*, *88*(3), 297–306.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, pp. 2554–2558. National Academy of Sciences.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. In *Proceedings of the National Academy of Sciences*, pp. 81:3088–3092. National Academy of Sciences.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, *41*, 401–410.

Hull, D. L. (1988). *Science as a Process.* University of Chicago Press.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic and procedural tasks. *Psychological Review, 96*(2), 208–233.

Humphreys, M. S., & Burt, J. (1994). Patterns of facilitation and inhibition in episodic and lexical access: Common functional requirements or common mechanisms?. Unpublished manuscript.

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix and TODAM models. *Journal of Mathematical Psychology, 33*(1), 36–67.

Humphreys, M. S., Wiles, J., & Dennis, S. (in press). Data structures and access processes: A first approximation to a theory of human memory. Behavioral and Brain Sciences.

Humphreys, M. S., Wiles, J., & Bain, J. D. (1992). Direct access: Cues with separate histories. In Meyer, D. E., & Kornblum, S. (Eds.), *Attention and Performance XXIV: A Silver Jubilee*, pp. 489–508. MIT Press, Cambridge, MA.

Jacoby, L. L. (1983). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory and Cognition, 9*, 21–38.

Jordan, M. I. (in press). Serial order: A parallel distributed processing approach. In Elman, J. L., & Rumelhart, D. E. (Eds.), *Advances in Connectionist Theory: Speech.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Jung, J. (1962). Transfer of training as a function of degree of first-list learning. *Journal of Verbal Learning and Verbal Behavior, 1*, 197–199.

Kail, R. (1984). *The Development of Memory in Children.* W H Freeman and Company, NY.

Karmiloff-Smith, A. (1992a). Abnormal phenotypes and the challenges they pose to connectionist models of development. Technical Report PDP.CNS.92.7, Carnegie Mellon University, Pittsburgh, PA.

Karmiloff-Smith, A. (1992b). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. MIT Press/Bradford Books, Cambridge, MA.

Karmiloff-Smith, A. (1992c). Nature, nurture and PDP: Preposterous developmental postulates?. *Connection Science*, *4*(3 & 4), 253–269.

Keppel, G. (1968). Verbal learning and memory. *Annual Review of Psychology*, *19*, 169–202.

Keppel, G., & Postman, L. (1966). Studies of learning to learn: III. Conditions of improvement in successive transfer tasks. *Journal of Verbal Learning and Verbal Behavior*, *5*, 260–267.

Kinsbourne, M., & George, J. (1974). The mechanism of the word frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 63–69.

Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review*, *85*, 363–394.

Koltz, S., & Johnson, N. L. (1982). *Encyclopedia of Statistical Sciences*. John Wiley and Sons, NY.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lewandowsky, S. (1991). Gradual unlearning and catastrophic interference: A comparison of distributed architectures. In Hockley, W. E., & Lewandowsky, S. (Eds.), *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*, pp. 445–476. Lawrence Erlbaum Associates, Hillsdale, NJ.

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review*, *96*(1), 25–57.

MacWhinney, B. (1993). The (il)logical problem of language acquisition. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pp. 61–70 Hillsdale, NJ. Lawrence Erlbaum Associates.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*, 252–271.

Mandler, G., Goodman, G. O., & Wilkes-Gibbs, D. L. (1982). The word frequency paradox in recognition. *Memory and Cognition*, *10*, 33–42.

Marr, D. (1982). *Vision*. Freeman, San Francisco, CA.

Martin, E. (1965). Transfer of verbal paired associates. *Psychological Review*, *72*, 327–343.

Martin, R. B., Simon, S., & Ditrichs, R. (1966). Verbal paired-associates transfer as a function of practice and paradigm shift. *Psychonomic Science*, *4* (12), 419–420.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In Bower, G. H. (Ed.), *The Psychology of Learning and Motivation*, pp. 109–165. Academic Press, NY.

McCormack, P. D., & Swenson, A. L. (1972). Recognition memory for common and rare words. *Journal of Experimental Psychology*, *95*, 72–77.

McCullers, J. C. (1965). Type of associative interference as a factor in verbal paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, *4*, 12–16.

McGeoch, J. A., & Irion, A. L. (1952). *The Psychology of Human Learning*. Longmans Green, New York.

McVaugh, W. H., & Postman, L. (1969). Studies of learning to learn: IX. Temporal changes in nonspecific transfer. *Journal of Verbal Learning and Verbal Behavior*, *8*, 745–753.

Melcher, N. (1993). Original representaion: Meaning in minds and machines. Unpublished.

Melton, A. W. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *2*, 1–21.

Moody, J. (1992). The effective number of parameters. In Moody, J. E., Hanson, S. J., & Lippmann, R. P. (Eds.), *Advances in Neural Information Systems 4*, pp. 847–854 San Mateo, CA. Morgan Kaufmann Publishers.

Morris, C. C., Bransford, J. D., & Franks, J. J. (1977). Levels of processing transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Mozer, M. C. (1992). Induction of multiscale temporal structure. In Moody, J. E., Hanson, S. J., & Lippmann, R. P. (Eds.), *Advances in Neural Information Processing Systems 4*, pp. 325–332 San Mateo: CA. Morgan Kaufmann.

Mozer, M. C. (1993). Neural net architecures for temporal sequence processing. To appear in: A. Weigend & N. Gershenfeld (Eds.) *Predicting the future and understanding the past*. Redwood City, CA: Addison-Wesley Publishing.

Murdock, B. B. (1962). The serial-position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609–626.

Murdock, B. B. (1987). Learning in a distributed memory model. In Izawa, C. (Ed.), *Current Issues in Cognitive Processes*, chap. 4, pp. 69–106. Lawrence Erlbaum Associates. The Tulane Flowerree Symposium on Cognition.

Murdock, B. B., & Kahana, M. J. (1992). An analysis of the list-strength effect. Submitted to the Journal of Experimental Psychology: Learning, Memory and Cognition.

Murdock, B. B., & Lamon, M. (1988). The replacement effect: Repeating some items while replacing others. *Memory & Cognition*, *16*(2), 91–101.

Murnane, K., & Shiffrin, R. M. (1991a). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *17*, 855–874.

Murnane, K., & Shiffrin, R. M. (1991b). Word repetitions in sentence recognition. *Memory and Cognition*, *19*(2), 119–130.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*, 135–183.

Nobel, P. A., & Shiffrin, R. M. (1992). Constraints on models of recognition and recall imposed by data on the time course of retrieval. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, pp. 1014–1019 Hillsdale, NJ. Lawrence Erlbaum Associates.

Nolfi, S., Parisi, D., Vallar, G., & Burani, C. (1990). Recall of sequences of items by a neural network. In Touretsky, D. S., Elman, J. L., Sejnowski, T. J., & Hinton, G. E. (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*. Morgan Kaufmann, San Mateo, CA.

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review*, *56*, 132–143.

Phillips, S. (1991). Serial recall using an Elman net with hints. Unpublished manuscript.

Phillips, S., & Wiles, J. (1993). Exponential generalizations from a polynomial number of examples in a combinatorial domain. To appear in the International Joint Conference on Neural Networks.

Piaget, J. (1970a). *Genetic Epistemology*. Columbia University Press, New York.

Piaget, J. (1970b). *Structuralism*. Basic Books, New York.

Piaget, J. (1971). *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes*. University of Chicago Press, Chicago.

Piaget, J., & Inhelder, B. (1973). *Memory and Intelligence*. Routledge and Kegan Paul, London.

Piatelli-Palmarini, M. (1989). Evolution, selection, and cognition: From " learning" to parameter setting in biology and the study of language. *Cognition*, *31*(1-44).

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review*, *91*(3), 281–293.

Plate, T. (1991). Holographic reduced representations: Convolution algebra for compositional distributed representations. In Mylopoulos, J., & Reiter, R. (Eds.), *Proceedings of the Twelfth International Conference on Artificial Intelligence*, pp. 30–35 Sydney, Australia. Morgan Kaufmann Publishers.

Plate, T. (1992). Holographic recurrent networks. In Giles, C. L., Hanson, S. J., & Cowan, J. D. (Eds.), *Advances in Information Processing Systems*, Vol. 5, pp. 34–41 San Mateo, CA. Morgan Kaufmann.

Plunkett, K., & Sinha, C. G. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology*, *10*, 209–254.

Pollack, J. B. (1988). Recursive Auto-Associative Memory: Devising compositional distributed representations. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, pp. 33–39 Montreal.

Postman, L. (1964). Studies of learning to learn: II. Changes in transfer as a function of practice. *Journal of Verbal Learning and Verbal Behavior*, *3*, 437–447.

Postman, L. (1968a). Studies of learning to learn: VI. General transfer in three-stage mediation. *Journal of Verbal Learning and Verbal Behavior*, *7*, 659–664.

Postman, L. (1968b). Studies of learning to learn: VIII. The influence of massive amounts of training upon the learning and retention of paired associate lists. *Journal of Verbal Learning and Verbal Behavior*, *7*, 790–796.

Postman, L. (1969). Experimental analysis of learning to learn. In Bower, G. H., & Spence, J. T. (Eds.), *Psychology of Learning and Motivation*, Vol. 3, pp. 241–297. Academic Press.

Postman, L. (1975). Verbal learning and memory. *Annual review of psychology*, *26*, 291–335.

Postman, L., Burns, S., & Hasher, L. (1970). Studies of learning to learn. X: Non-specific transfer effects in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *9*, 707–715.

Postman, L., Keppel, G., & Zacks, R. (1968). Studies of learning to learn: VII. The effects of practice on response integration. *Journal of Verbal Learning and Verbal Behavior*, *7*, 776–784.

Postman, L., & Schwartz, M. (1964). Studies of learning to learn: I. Transfer as a function of method of practice and class of verbal materials. *Journal of Verbal Learning and Verbal Behavior*, *3*, 37–49.

Postman, L., & Stark, K. (1967). Studies of learning to learn: IV. Transfer from serial to paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, *6*, 339–353.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of Associative Memory. *Psychological Review*, *88*, 93–134.

Rao, K. V. (1983). The word frequency effect in situational frequency estimation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *9*, 73–81.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*(2), 285–308.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*, 163–178.

Regier, T. (1992). The acquisition of lexical semantics for spatial terms: A connectionist model of perceptual categorization. Technical Report TR-92-062, International Computer Science Institute, Berkeley, CA.

Reilly, R. (1993). A connectionist attentional shift model of eye-movement control in reading. In *Proceedings of the 15th Annual Meeting of the Cognitive Science Society* Boulder, CO. To appear.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypothesis. *Journal of Experimental Psychology*, *92*, 365–372.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., & McClelland, J. L. (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, pp. 318–362. MIT Press, Cambridge, MA.

Schmidhuber, J. (1991). Neural sequence chunkers. Technical Report FKI-148-91, Technische Universitaet Muenchen, Institut fuer Informatik, Munich, Germany.

Schmidhuber, J. (1992). A fixed size $O(n^3)$ time complexity learning algorithm for fully recurrent networks. *Neural Computation*, *4*(2), 243–248.

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science*, *9*, 211–212.

Schulman, A. I. (1976). Memory for rare words previously rated for familiarity. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 301–307.

Schulman, A. I., & Lovelace, E. A. (1970). Recognition memory for words presented at slow or rapid rate. *Psychonomic Science*, *21*, 99–100.

Schwenn, E., & Postman, L. (1967). Studies of learning to learn: V. Gains in performance as a function of warm-up and associative practice. *Journal of Verbal Learning and Verbal Behavior*, *6*, 565–573.

Shepard, R. N. (1967). Recognition memory for words, sentences and pictures. *Journal of Verbal Learning and Verbal Behaviour*, *6*, 156–163.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological systems. *Science*, *237*, 1317–1323.

Shiffrin, R. M., & Murnane, K. (1991). Composition, distribution, and interference in memory. In Hockley, W. E., & Lewandowsky, S. (Eds.), *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*, pp. 331–347. Lawrence Erlbaum Associates, Hillsdale, NJ.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). The list-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*, 179–195.

Shultz, T. R. (1991). Simulating stages of human cognitive development with connectionist models. In Birnbaum, L., & Collins, G. (Eds.), *Machine Learning: Proceedings of the Eighth International Workshop*, pp. 105–109. Morgan Kaufman.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1–23.

Smolensky, P., Legendre, G., & Miyata, Y. (1992). Principles for an integrated connectionist/symbolic theory of higher cognition. Technical Report CU-CS-600-92, University of Colorado.

Spence, J. T., & Lair, C. V. (1965). Associative interference in the paired-associate learning of remitted and nonremitted schizophrenics. *Journal of Abnormal Psychology*, *70*(2), 119–122.

Stornetta, W. S., & Huberman, B. A. (1987). An improved three-layer back-propagation algorithm. In Caudill, M., & Butler, C. (Eds.), *IEEE First International Conference on Neural Networks*, pp. 637–645 San Diego, CA. IEEE.

Sumby, W. H. (1963). Word frequency and serial position effects. *Journal of Verbal Learning and Verbal Behavior*, *1*, 443–450.

Swanson, J. M., & Wickens, D. D. (1970). Preprocessing on the basis of frequency of occurrence. *Quarterly Journal of Experimental Psychology*, *22*, 378–383.

Swets, J. A., & Green, D. M. (1961). Sequential observations by human observers of signals in noise. In Cherry, C. (Ed.), *Information Theory: Proceedings of the fourth London symposium*, pp. 177–195 London. Butterworth.

Thorndike, E. L. (1965). *Animal Intelligence: Experimental Studies.* Hafner Publishing Company, New York, NY. Original published in 1911 by Teachers College, Columbia University.

Touretzky, D. S., & Pomerleau, D. A. (1991). Neural networks for real-world problems. Tutorial given at 12th International Joint Conference on Artificial Intelligence, Sydney, Australia.

Tulving, E. (1983). *Elements of Episodic Memory.* Oxford Psychology Series No. 2. Clarendon Press, Oxford.

Tulving, E., & Hastie, E. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology*, *92*(3), 297–304.

Tulving, E., & Madigan, S. A. (1970). Memory and verbal learning. *Annual Review of Psychology*, *21*, 437–484.

Tulving, E., McNulty, J. A., & Ozier, M. (1965). Vividness of words and learning to learn in free-recall learning. *Canadian Journal of Psychology*, *19*, 242–252.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity processes and retrieval processes in episodic memory. *Psychological Review*, *80*, 352–373.

Underwood, B. J. (1972). Word recognition memory and frequency information. *Journal of Experimental Psychology*, *94*, 276–283.

Underwood, B. J., & Freund, J. J. (1970). Testing effects in the recognition of words. *Journal of Verbal Learning and Verbal Behavior*, *9*, 117–125.

Underwood, B. J., & Schulz, R. W. (1960). *Meaningfulness and verbal learning.* Lippincott, Philadelphia.

Wagner, D. A. (1978). Memories of Morocco: The influence of age, schooling, and environment on memory. *Cognitive Psychology*, *10*, 1–28.

Waibel, A. (1989). Modular construction of time-delay neural networks for speech recognition. *Neural Computation*, *1*(1), 39–46.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. (1987). Phoneme recognition using time-delay neural networks. Technical Report TR-1-0006, ATR Interpreting Telephony Research Laboratories, Japan.

Watkins, M. J. (1990). Mediationism and the obfuscation of memory. *American Psychologist*, *45*, 328–335.

Wiles, J., & Bloesch, A. (1992). Operators and curried functions: Training and analysis of simple recurrent networks. In Moody, J. E., Hanson, S. J., & Lippmann, R. P. (Eds.), *Advances in Neural Information Processing Systems 4*, pp. 325–332 San Mateo: CA. Morgan Kaufmann.

Wiles, J., Humphreys, M. S., Bain, J. D., & Dennis, S. J. (1990). Direct memory access using two cues: Finding the intersection of sets in a connectionist model. Technical Report 186, Department of Computer Science, The University of Queensland, Australia, 4072.

Wiles, J., & Phillips, S. (1991). Serial recall of binary sequences. Unpublished manuscript.

Williams, R. J., & Zipser, D. (1989). Experimental analysis of the real-time recurrent learning algorithm. *Connection Science*, *1*(1), 87–111.

Williams, R. J., & Zipser, D. (1990). Gradient-based learning algorithms for recurrent networks. In Chauvin, Y., & Rumelhart, D. E. (Eds.), *Backpropagation: Theory, Architectures and Applications*, pp. 1–42. Erlbaum, Hillsdale, NJ.

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*(2), 345–355.

Zechmeister, E. B. (1969). Orthographic distinctiveness. *Journal of Verbal Learning and Verbal Behavior*, *8*, 754–761.

Zechmeister, E. B. (1972). Orthographic distinctiveness as a variable in word recognition. *American Journal of Psychology*, *85*, 425–430.

Zechmeister, E. B., Curt, C., & Sebastian, J. A. (1978). Errors in a recognition memory task are a U-shaped function of word frequency. *Bulletin of the Psychonomic Society*, *11*, 371–373.

Appendix A

Learning and Meaning

In order to appreciate the importance of learning models it is first necessary to outline the physical symbol system hypothesis (Newell, 1980), which underpins cognitive research in the information processing paradigm, and to retrace a dispute which has raged in the literature for well over a decade, but which has more recently been dubbed the *symbol grounding problem* (Harnad, 1990). The purpose of this discussion is to bring into focus the representational assumptions on which rational analysis makes its foundations, and to illustrate the difficulties this creates for an explanation of learning. *Interactivism* (Bickhard, 1991) is then presented as an alternative representational scheme that facilitates the integration of learning into cognitive systems.

What is the physical symbol system hypothesis?

The physical symbol system hypothesis is the paradigm under which research in artificial intelligence has proceeded and is considered by its proponents to underpin cognitive science. Newell (1980) states:

> The physical symbol system is to our enterprise what the theory of evolution is to all biology, the cell doctrine to cellular biology, the notion of germs to the scientific study of disease, the notion of tectonic plates to structural geology.

(p. 136, Newell, 1980)

Stated simply it says:

> The necessary and sufficient condition for a physical system to exhibit general intelligent action is that it be a physical symbol system.

(p. 170, Newell, 1980)

where the critical aspects of a symbol system for our discussion are (adapted from Harnad, 1990):

(1) There exist arbitrary, discrete, atomic, physical tokens called symbols.

(2) These symbols are manipulated on the basis of explicit rules that combine and recombine symbols in a concatenative fashion forming expressions.

(3) These rules are also expressions.

(4) All expressions of the system are semantically interpretable.[1]

Recently, however, the plausibility of the hypothesis has come under question (Harnad, 1990; Bickhard, 1991). The major difficulty involves the fourth point from the list above, namely, how the symbols acquire their meaning. The next section deals with what is known as the symbol grounding problem.

The Symbol Grounding Problem

Fodor's Anti-Constructivist Arguments: In a debate with Piaget concerning language and learning, Fodor (1980) said:

> ... it seems to me that there is a sense in which there isn't any theory of learning, ... I'll argue not only that there is no learning theory but that in certain senses there certainly couldn't be; the very idea of concept learning is, I think confused. ... The reason people think that there is such a theory is that they confuse a theory of concept acquisition with a theory that has a quite different kind of logical structure, what I will call a theory of "fixation of belief." Although there are no theories of concept learning, there are theories of fixation of belief, and I think they are very deep and very important. The trouble is that to get any of them to work you have to be radically innatist about the origin of concepts. (p. 141)

[1] The fourth aspect specifies that all of the expressions that occur in the system can *systematically* be assigned a meaning.

When formulating this argument, Fodor (1980) refers to *concept learning* experiments of the following form. A subject is asked to stack into two piles a set of cards with which she is presented. These cards display red circles, green triangles, blue tetrahedrons and so on. As each card is stacked the experimenter tells the subject whether the card has been stacked "correctly". The rule that the experimenter is following might be that all red circles and green triangles should be placed in one pile and everything else in the other pile.

The point that Fodor (1980) emphasizes is that the concepts that are being learned are simple combinations of the basic concepts already available to the subjects. Nothing new is coming into existence since the hypotheses that became fixed during the "learning" were already present in a sense. What is being acquired is a mapping from the possible expressions of the system to their truth values. Fodor (1980) argues that the term concept learning should be reserved for the process during which the set of representable concepts is enlarged. He makes an analogy between Piagetian stages and logics, and argues that statements such as "children at stage i don't have transitivity" implies that children at stage i are using a weaker logic than their stage i+1 counterparts. Learning is the process by which humans come to instantiate stronger logics. This concept of learning is much stronger than that of fixation of belief, but is necessary to escape radical nativism, as highlighted by the following example.

Symbols, Meaning and Chinese in a Vacuum: Harnad (1990) asks us to consider the task of learning Chinese as a first language purely from a Chinese/Chinese dictionary. One would open to the first page and choose a word. The only way in which they could decipher the words by which its meaning was expressed would be to make their way to yet another page. If one were to ever find the next word, given that the organisation of the dictionary would also be completely obscure at this point, one would simply be redirected to yet another set of words. The regress would never stop and, hence, the meaning could not be established. It is for this reason that Fodor (1980) supports the nativist position. It is certainly possible for us to put together basic concepts in order to form *phrasal* concepts. There is, however, no way for us to "learn" the meanings of basic concepts, hence, they must already exist, that is, they must be innate. The result of our long evolutionary history is

that the world is connected to the mind so that the basic symbols have intrinsic meaning.

The Essence of the Problem - Encodingism

Bickhard (1991) points out that this explanation does not avoid the symbol grounding regress. If the only sort of learning that can occur is what Fodor (1980) refers to as "belief fixation" then it would seem that this should be as true of evolution as it is of what ever mechanism the mind might use in order to learn. The difficulty is logical and not contingent on the particular mechanism.

The argument then becomes more serious. The nativist position is already difficult methodologically, since the major explanatory weight is being attributed to a mechanism that can not be observed directly and the modeller must make a large number of somewhat arbitrary assumptions in order to predict the subject's behaviour. That evolution can do no better, however, implies that such a mechanism simply could not have come into being. So where is the difficulty?

The problem lies in the way in which the symbols receive their meaning (Bickhard, 1991)[2]. The representational system that is being used is a symbol system (or an *encodingism* in Bickhard's terminology). The only way in which an expression receives meaning is by parasitically acquiring it from the other expressions of which it is composed. The interpreter reads the symbol string and imposes the meaning that forms in their heads back onto the symbol string. While composition is the only mechanism by which meaning is attained there is no logical place to halt the regress, and no way for meaning to be grounded.

The fact that the expressions being used are "semantically interpretable" promotes the illusion that they have intrinsic meaning. In essence, the system is receiving a free lunch from the reader's interpretive machinery.

A Solution to the Symbol Grounding Problem - Interactivism

To address the encoding problem, Bickhard (1991) proposes *interactivism* as an alternative method for determining the meaning of representations. In an encodingism,

[2]Bickhard (1991) uses the term "representational content" rather than "meaning". A state is said to have *representational content* when that state is used to generate an appropriate behaviour, that is, it contributes to the systems ability to successfully interact.

the process of forming a representation is identical to the process of determining the meaning of a representation. As a consequence, the only items that can enter into forming a representation are other representations all of which already have a meaning and there is no escape from the regress. The first step in building an interactive system is to separate the condition of being a representation from that of having meaning (Bickhard, 1991). Interactive systems allow a representation to form and gain its meaning either at a later time or as the culmination of a gradual process.[3]

As Bickhard (1991) points out, divorcing the notion of "representation" from that of "meaning" requires reconceptualising both of these constructs. What is required is a "representation" that does not carry meaning and "meaning" that can come into being subsequent to the establishment of the representation.

Bickhard's redescription of "representation" and "meaning" focusses on the interaction between the knowing system and the environment. Such a system has internal states that occur as a consequence of previous internal states and the current environmental stimulation. Furthermore, there will be a set of environments that lead to a given internal state and this internal state can be said to *differentiate* its set of environments from other sets of environments that result in the system attaining different internal states. These states are Bickhard's representations. They are more primitive than the usual concept of a representation in that they do not represent anything. They are correlated with their environmental sets, but they do not as yet have meaning.

In interactivism, "meaning" is determined by successful interaction with the environment. Suppose that there are two internal states A and B which occur in the knowing system as a consequence of environments X and Y. Suppose further that when the knowing system is in state A it employs procedure P133. When the system is in state B it employs procedure P47. Initially, the connection of internal states to procedures might be quite arbitrary. As the system interacts with its environment it alters the procedures that it employs so that they become appropriate to the goals of the system. As it makes these alterations, however, it is gaining knowledge, that

[3]Recall that the fourth aspect of a symbol system outlined above was that all of the expressions must be semantically interpretable. It is this condition that must be relaxed in order to escape the regress.

is, the representations are gaining meaning[4]. Hence, Bickhard develops a notion of representations that can occur without meaning, and meaning that emerges as a consequence of interaction.

What it means for the system to have a goal requires some clarification. As Melcher (1993) points out, it is not sufficient to think of a goal as a meaningful representation in the knowing system, as is often done in cognitive science and artificial intelligence. Since the notion of goal is being used to define "meaning", the notion of meaning cannot be presupposed. Melcher (1993) defines a goal in behavioural terms. A system is said to have the goal to attain C in an environment X with current state A, if it continues to execute a given procedure, in this case P133, until C is attained and then stops. The knowing system may execute other procedures at the same time and in response to the execution of P133. The critical point is that P133 continue to be executed until C is attained and not thereafter. Such a definition avoids presupposing meaningful representation to define goals, while providing a notion of what error means for the system (i.e. if the system fails to attain C it is in error).

Notice is that this definition of meaning is very different from that used in a symbol system. The way in which meaning is attributed is not dependent on the way in which an expression is composed. Instead, meaning accrues as the consequence of interaction with the environment. Consequently, the encodings regress is circumvented, and Fodor (1980)'s strong nativist conclusions are avoided.

[4]This example is due to Melcher (1993).

Appendix B

Example of Environmental Data Processing

This appendix provides an example of the text processing that was undertaken when investigating the environment relevant to the list strength effect (see chapter two). Section B.1 shows a sample article from the online version of the University of Minnesota Daily. Section B.2 shows the same article once noise components such as punctuation, numerals and functor words had been removed. Section B.3 demonstrates how each article was divided into study and test segments.

B.1 Raw Text Sample

Publish Date: 03/01/1991

Headline: Baseball has made a permanent impression on the way I

March is the time of year when political columnists like my personal hero, George Will, take a break from the bump-and-grind of international affairs and domestic hanky-panky to wax philosophical about professional baseball. I can only hope to live up to such greatness. In a recent national poll, Americans voted football as their favorite sport by a wide margin. Baseball brooded uncomfortably in seventh place, sandwiched between rhythmic synchronized ice ballet and cockfighting. How can the decline of baseball, once called The Great American Pastime, be explained? Some say skyrocketing salaries are to blame. The Oakland Athletics, for instance,

have 13 players who make more than $1 million per year. Five of those players will be making more than $3 million this year. Another theory says that baseball is too slow for our modern times, where wars last three days and puberty is over and done with on a weekend. A third theory places the blame entirely on the shoulders of Pete Rose. The all-time hit leader has thrilled many a young autograph-seeker with his winning style. Maybe it's his special way of tousling a boy's hair, reaching out his hand ever so gently and saying, "Five bucks, sonny. Five bucks." But enough sentiment. It's March, spring training has begun, and hearts across America are fluttering like a knuckleball as the fields of dreams come to life with the sights and smells of baseball ...

HOW BASEBALL IMITATES THE UNIVERSE

Baseball is poetry, a nine-act play, a crotch-scratching operetta. The beauty of baseball is that, just like our universe, there is no clock. Well, actually there is a clock, usually several. The clocks are there so that fans can see how long this game without a clock is lasting and if they'll be able to make it home in time for the news. In theory, a game could last forever, although the players might get a little bushed. There are no boundaries in baseball. In theory, a ball could be hit an infinite distance. According to Einstein, an outfielder chasing such a ball would return in the past. How far in the past would depend on how good a jump he got on the ball. Perhaps someday an astonished paleontologist will unearth a skeleton wearing a Twins uniform from the La Brea tar pits. Also like the universe, if you look up at a baseball game, you'll see a bunch of lights.

ZEN AND THE ART OF ROD CAREW

Rod Carew sits in the dugout at his hitting school in southern California, watching the steady drizzle make Rorschach prints on the infield. He sits as motionless as a stunned carp, the wad in his cheek beautiful and delicate. "I can pick out any object," the seven-time batting champion says. "Your shoe, for instance. Then I concentrate on it until I'm completely at peace, like I'm drifting in a giant soap bubble. I think about the fact that there are no clocks in baseball and the boundaries extend into infinity and then I take a deep breath and ..." The tobacco juice soars majestically out of his mouth and scores a direct hit on my shoe. It sits there motionless, as if even after releasing it the Master still has control.

"See what I mean?" he says.

"Yes, I do."

I think about Opening Day and smile.

WON'T SOMEBODY PLEASE TALK TO ME?

Conversation in baseball is like a wicked curveball: You never know where it's going. If you ride the buses with minor-leaguers, teams like the Iowa City Bandits or the Riverside Rangers, you hear the lore of the game, stories passed on from generation to generation.

"... so then I said to her, Come here and I'll #$*&%*! your *&$#%*!" "Har, har," the boys of summer laughed.

"That's positively Homeric," I said.

"Shaddup and get us a beer!"

BIG KIDS IN THEIR PAJAMAS

In the chaos of the universe, baseball is an oasis of Homeric sanity. In the shady quietude of backyards and parks, the game is passed down from fathers to their sons in a transcendental genetic home run, a living allegory of all that has gone before and all that is still to come. On warm summer weekends these fathers take their sons to see the big-leaguers.

Where did you go?

Inside.

What did you do?

Baseball.

Our hearts leap as we catch the first sight of the giant trash bag in right field. The smell of burning nachos, the thrill of the fraying carpet. My boy asks why our seats along the third-base line are facing the outfield instead of home plate. "Because that's where the balls are hit," I tell him. Baseball, in essence, is a children's game. You play catch even though you throw like a girl, swing at the ball and miss, then grow up and write a newspaper column.

Baseball, in essence, is a thinking-man's game.

Baseball, in essence is ...

"Look out, Pop, look out!"

THUNK!

As the ushers carried me to the first aid station, I traced the stitch marks imprinted on my forehead by the foul ball. I smiled. At last, I had become One with the Game.

B.2 Clean Text Sample

publish date headline baseball made permanent impression march time year political columnists personal hero george take break bump grind international affairs domestic hanky panky wax philosophical professional baseball hope live greatness recent national poll americans voted football favorite sport wide margin baseball brooded uncomfortably seventh place sandwiched rhythmic synchronized ice ballet cockfighting decline baseball once called great american pastime explained say skyrocketing salaries blame oakland athletics instance players million per year five players making million year another theory says baseball too slow modern times wars last three days puberty weekend third theory places blame entirely shoulders pete rose time hit leader thrilled young autograph seeker winning style maybe special tousling boy hair reaching out hand gently saying five bucks sonny five bucks enough sentiment march spring training begun hearts across america fluttering knuckleball fields dreams life sights smells baseball baseball imitates universe baseball poetry nine act play crotch scratching operetta beauty baseball just universe clock actually clock usually several clocks fans see long game clock lasting ll able home time news theory game last forever although players might get little bushed boundaries baseball theory ball hit infinite distance according einstein outfielder chasing ball return past far past depend good jump got ball perhaps someday astonished paleontologist unearth skeleton wearing twins uniform la brea tar pits universe look baseball game ll see bunch lights zen art rod carew rod carew sits dugout hitting school southern california watching steady drizzle rorschach prints infield sits motionless stunned carp wad cheek beautiful delicate pick out object seven time batting champion says shoe instance concentrate until completely peace drifting giant soap bubble think fact clocks baseball boundaries extend infinity take deep breath tobacco juice soars majestically out mouth scores direct hit shoe sits motionless releasing master still

control see mean says yes think opening day smile won somebody please talk conversation baseball wicked curveball never going ride buses minor leaguers teams iowa city bandits riverside rangers hear lore game stories passed generation generation said her ll har har boys summer laughed positively homeric said shaddup get beer big kids pajamas chaos universe baseball oasis homeric sanity shady quietude backyards parks game passed down fathers sons transcendental genetic home run living allegory gone still warm summer weekends fathers take sons see big leaguers did inside did baseball hearts leap catch first sight giant trash bag right field smell burning nachos thrill fraying carpet boy asks seats along third base line facing outfield instead home plate balls hit tell baseball essence children game play catch though throw girl swing ball miss grow write newspaper column baseball essence thinking man game baseball essence look out pop look out thunk ushers carried first aid station traced stitch marks imprinted forehead foul ball smiled last become one game

B.3 The Segments Extracted from the Clean Text Sample

Study Segment

publish date headline baseball made permanent impression march time year political columnists personal hero george take break bump grind international affairs domestic hanky panky wax philosophical professional baseball hope live greatness recent national poll americans voted football favorite sport wide margin baseball brooded uncomfortably seventh place sandwiched rhythmic synchronized ice ballet cockfighting decline baseball once called great american pastime explained say skyrocketing salaries blame oakland athletics instance players million per year five players making million year another theory says baseball too slow modern times wars last three days puberty weekend third theory places blame entirely shoulders pete rose time hit leader thrilled young autograph seeker winning style maybe special tousling boy hair reaching out hand gently saying five bucks sonny five bucks enough sentiment march spring training begun hearts across america fluttering knuckleball fields dreams

Test Segment

life sights smells baseball baseball imitates universe baseball poetry nine act play crotch scratching operetta beauty baseball just universe clock actually clock usually several clocks fans see long game clock lasting ll able home time news theory game last forever although players might get little bushed boundaries baseball theory ball hit infinite distance according einstein outfielder chasing ball return past far past depend good jump got ball perhaps someday astonished paleontologist unearth skeleton wearing twins uniform la brea tar pits universe look baseball game ll see bunch lights zen art rod carew rod carew sits dugout hitting school southern california watching steady drizzle rorschach prints infield sits motionless stunned carp wad cheek beautiful delicate pick out object seven time batting

Appendix C

Regression Analysis of the List Strength Environmental Data

Table C.1: Significance levels of adding each of the independent variables on the prediction of recurrence probability for each of 20 replications.

N	General Repetition Rate	Target Repetition Rate	Word Frequency
198	0.2743	* 0.0000	0.7380
198	0.7529	* 0.0000	0.5329
198	0.3420	* 0.0000	0.1048
198	0.6314	* 0.0000	* 0.0007
197	0.7778	* 0.0000	0.0528
196	* 0.0295	* 0.0000	* 0.0031
195	0.7242	* 0.0000	* 0.0000
190	0.3640	* 0.0000	* 0.0000
189	0.3312	* 0.0000	* 0.0000
186	0.9261	* 0.0004	* 0.0000
181	0.2165	* 0.0002	* 0.0000
173	0.0925	* 0.0032	0.5342
163	0.9881	0.0694	* 0.0003
151	0.5015	* 0.0000	* 0.0000
138	0.3783	* 0.0001	* 0.0032
129	0.5412	* 0.0190	* 0.0000
120	0.4079	0.2661	* 0.0000
107	0.7930	0.7128	* 0.0012
96	0.8204	* 0.0185	0.2114
81	0.1329	* 0.0000	* 0.0000

Note. An asterisk indicates a significant result (i.e. $p < 0.05$). N is the number of observations in that sample.

Appendix D

Rational Analysis of Selected Phenomena

Anderson's *rational analysis*[1] was the first attempt within the memory literature to characterise the affect of the environment on the performance characteristics of the memory system. This appendix provides rational analyses for the phenomena that were discussed in chapter two, namely, the word frequency (section D.1) and list strength effects (section D.2) in recognition, and the ABABr (section D.3) and crossed associates (section D.4) cued recall paradigms.

D.1 The Word Frequency Effect in Recognition

Anderson and Milson (1989) explain the word frequency effect[2] in terms of the context factor (i.e. the cue set to memory trace probability). The critical ratio in their analysis is $\frac{n(word|word\ in\ cue set)}{n(word)}$. The numerator is the probability that the the memory trace "word" is needed given that "word" is present as a cue. The denominator is the probability that the memory trace word is needed. They assume that the numerator is fairly constant for all words and that performance will be inversely proportional to the word frequency due to the denominator.

There is, however, a point of confusion in the argument presented by Anderson and Milson (1989). When examining the context factor, they refer to the Bayesian

[1] See chapter one, section 5.1 for an outline of rational analysis.

[2] See chapter two, section one for an explanation of the word frequency effect in recognition.

odds formula. There are, however, two reasons why considering the odds formula rather than the probability formula seems curious. Firstly, in the consideration of the history factor they looked at the probability formula not the odds formula. Their factor in this case was $P(A|H_A)$ as opposed to $\frac{P(A|H_A)}{P(\bar{A}|H_A)}$. Secondly, in the initial formulation of the optimisation problem they require $P(A)$, not the odds factor. Their concentration on the odds formula is important because it has consequences for the argument that followed. The claim is made that $P(i)$ and $P(i|\bar{A})$ would be nearly identical because conditionalizing on the nonrelevance of one memory structure out of such a large number of possibilities cannot change the probability of any cue. In general, this may well be the case. Consider, however, the recognition paradigm. Here $P(word)$ must be compared with $P(word|word\ not\ in\ trace)$. It seems entirely possible that the fact that the word is not in the trace will affect the probability of the same word being in the cue set. The argument that they make is not undermined, however, since had Anderson and Milson (1989) used the probability formula rather than the odds formula the above objection would have been circumvented.

The environmental analysis carried out in chapter two suggests that the history factor is also implicated in the word frequency effect. Anderson and Milson (1989) list four assumptions of the augmented Burrel model that they use to determine the history factor. They are:

(a) Items vary in their desirability, which controls the rate at which they are accessed.

(b) The rate of use of an item decays with time.

(c) Items vary in their rates of decay.

(d) Items undergo revivals of interest.

The environmental analysis in chapter two demonstrates that:

$$P(HighFreqWord|H_{HighFreqWord}) < P(LowFreqWord|H_{LowFreqWord})$$

for the duration of the experimental context. That is, immediately following the presentation of a word its recurrence (or need) probability will be higher if it is a low frequency word. Given that the overall probability of a low frequency word is by definition lower than that for the high frequency word, the decay rate for low

frequency words must be greater than that for high frequency words. Hence, if assumption (c) is parameterised with the frequency of the word, an account for the qualitative affects of frequency can be constructed. Anderson and Milson (1989) suppose that the decay rates are exponentially distributed (i.e. $\alpha e^{-\alpha d}$ where d is the decay rate and $1/\alpha$ is the mean of the distribution). Hence, the short lived "flashes-in-the-pan" items will be frequent while the "classics" will occur rarely. To account for the results above, α could be replaced with $\alpha * frequency$[3]. Hence, the mean decay rate for a high frequency word will be less than that for a low frequency word and low frequency words will start with greater need probabilities and decay more quickly (see figure 2.5).

D.2 The List Strength Effect in Recognition

The main effect of study set frequency has already been dealt with successfully in rational analysis by Anderson and Milson (1989). As yet, though, the rational model has not been applied to the list strength finding[4].

The analysis of the list strength effect (Shiffrin et al., 1990; Ratcliff, 1990) has centered around the ratio of ratios measure given by the following formula:

$$R_r = \frac{d'(mixed\,strong)/d'(mixed\,weak)}{d'(pure\,strong)/d'(pure\,weak)} \tag{D.1}$$

As a first approximation, Anderson and Milson (1989) assume that performance (i.e. d') is proportional to the conditional probability, $P(A|H_A\&Q)$. Since R_r is a ratio, the $P(A|H_A\&Q)$ terms can be substituted for the corresponding d' terms. The context factor should not alter as a function of the list strength manipulations. The important factor to analyse is the history factor. Anderson (1989) introduces a model of the history factor based on the work in non-human information retrieval systems (e.g. libraries) by Burrell. The extensions introduced in the augmented Burrell model (Anderson & Milson, 1989) are significant for distinguishing between blocked versus spaced presentation. The list strength effect has been demonstrated in both conditions (Ratcliff, 1990), so for simplicity, the original Burrell model will

[3]It is important to note that $\alpha * frequency$ parameterises the distribution of decay rates and is not the decay rate itself.

[4]See chapter two, section two for a description of the list strength effect.

be assumed for this analysis. The prediction of the history factor is given by:

$$P(A|H_A) = \frac{v+n}{b+t} \tag{D.2}$$

where v and b are parameters of the assumed gamma distribution, n is the number of presentations of the item and t is the total elapsed time since the first presentation of the item.

Murnane and Shiffrin (1991a) discuss the affect of time lag and note that it will produce an artifactual positive list strength effect in the mixed-pure paradigm, which is commonly used to determine the affect of strength. They note that when lag is controlled the results are more consistent than when it is not. This analysis follows Shiffrin et al. (1990)[5] in assuming that t is constant[6]. Substituting for each d' by the corresponding need probability produces:

$$R_r = \frac{[v + n_{strong}/b + t]/[v + n_{weak}/b + t]}{[v + n_{strong}/b + t]/[v + n_{weak}/b + t]} = 1 \tag{D.3}$$

where n_{weak} is the number of presentations of the weak items and n_{strong} is the number of presentations of the strong items.

Hence, on the basis of the need probabilities, rational analysis predicts no list strength effect. It is important, however, to consider how these need probabilities are mapped into probabilities and latencies of retrieval.

The simple monotonicity argument that Anderson and Milson (1989) invoke to link need probabilities to memory performance is satisfactory for accuracy but will not hold in the case of latency. The retrieval mechanism supposes that memory traces are assessed one at a time in order of their need probabilities until either the appropriate trace is found, or the need probability (subject to noise) falls below a threshold. Accuracy is dependent purely on the need probability of the item and the threshold. Hence, there will be no affect on accuracy of list strength manipulations. Interestingly, however, there should be an affect on latency. The inclusion of other strong items will potentially create traces with high need probabilities that must be rejected, thus increasing the latency[7].

[5] See Shiffrin et al. (1990)'s analysis of the TODAM (Murdock, 1982) model.

[6] The case explored in experiment 3 of Murnane and Shiffrin (1991a).

[7] This effect will depend critically on the parameters. While it is the case that $P(A|H_A)$ will be greater for the strong items, $\frac{P(i|A)}{P(i)}$ will favour the correct trace. The parameters will determine the impact of each factor.

Rational analysis is able to account for the affects of repetition in recognition. Both the affect of study set frequency (Anderson & Milson, 1989) and the lack of a list strength effect are consequences of the model. In addition, the model predicts that the latency of nonrepeated items will increase as the degree of repetition increases.

D.3 The ABABr Paradigm

The ABABr paradigm (Jung, 1962; Barnes & Underwood, 1959) involves two lists of pairs of words in which the second is the first list rearranged. In practice, subjects are able to respond well above chance on this task demonstrating that they do not rely on two way associations alone[8].

The first issue to be decided when applying rational analysis to the ABABr paradigm is what memory trace must be retrieved. Suppose, without lose of generality, that the lists are composed as follows:

List 1

Table Road

Grass Violin

List 2

Table Violin

Grass Road

One might think that the contenders would be the "Road" memory trace and the "Violin" memory trace. However, when one looks at the way in which Anderson and Milson (1989) deal with paired associate learning, the pairs of items are bundled into single memory traces. It is this binding of elements which is critical in being able to account for this task. In the ABABr context the three relevant aspects, that is, the list number, and the first and second words are considered to be bundled into single memory traces. After having been presented with the above lists, the traces that

[8]See chapter two, section three for an explanation of the paradigm and a recount of its implications for the order of associations in the memory system as outlined by Humphreys et al. (1989).

would have strong support from the history factor would be {List1, Table, Road}, {List1, Grass, Violin}, {List2, Table, Violin} and {List2, Grass, Road}.

What are the relevant need probabilities when the subject is presented with the cue set {List1, Table}? Equation 5 from Anderson and Milson (1989) expresses the need probability of a memory trace in terms of the history and context factors.

$$P(A|H_A\&Q) = P(A|H_A) * \prod_{i \in Q} \frac{P(i|A)}{P(i)} \tag{D.4}$$

where A is a memory trace, H_A is the history of that memory trace and Q is the set of cues that are currently available. The history factor determines what influence the history of the memory trace will have on its probability of need while the context factor determines the contribution of the cues that are presented to the subject[9].

Substituting the cues and the {List1, Table, Road} memory trace from the current example gives:

$$P(\{List1, Table, Road\}|H_{\{List1,Table,Road\}}\&\{List1, Table\}) = \tag{D.5}$$

$$P(\{List1, Table, Road\}|H_{\{List1,Table,Road\}}) * \prod_{i \in \{List1,Table\}} \frac{P(i|\{List1, Table, Road\})}{P(i)} \tag{D.6}$$

Expanding the cue set out:

$$P(\{List1, Table, Road\}|H_{\{List1,Table,Road\}}\&\{List1, Table\}) =$$

$$P(\{List1, Table, Road\}|H_{\{List1,Table,Road\}}) * \frac{P(List1|\{List1,Table,Road\})}{P(List1)} * \frac{P(Table|\{List1,Table,Road\})}{P(Table)} \tag{D.7}$$

[9]The use of the term context to denote the cue set is potentially confusing. Within the memory modelling literature, context often refers to the environmental cues only and does not normally include the study and test materials. For instance, Raaijmakers and Shiffrin (1981) refer to contextual information as the "temporal and situational factors that might also be present in STS at a given time, including environmental details, physical sensations, emotional feelings, and all thought processes not directly relevant to name production.".

Anderson and Milson (1989) further assume that:

$$\frac{P(i|A)}{P(i)} = \prod_{x \in A} \frac{P(i|x)}{P(i)} \tag{D.8}$$

where the x are the components of the memory trace.
Hence:

$$\begin{aligned} &P(\{List1, Table, Road\}|H_{\{List1,Table,Road\}}\&\{List1, Table\}) = \\ &\quad P(\{List1, Table, Road\}|H_{\{List1,Table,Road\}}) * \\ &\quad \frac{P(List1|List1)}{P(List1)} * \frac{P(List1|Table)}{P(List1)} * \frac{P(List1|Road)}{P(List1)} * \\ &\quad \frac{P(Table|List1)}{P(Table)} * \frac{P(Table|Table)}{P(Table)} * \frac{P(Table|Road)}{P(Table)} \end{aligned} \tag{D.9}$$

The need probability of the {List1, Grass, Violin} trace is given by:

$$\begin{aligned} &P(\{List1, Grass, Violin\}|H_{\{List1,Grass,Violin\}}\&\{List1, Table\}) = \\ &\quad P(\{List1, Grass, Violin\}|H_{\{List1,Grass,Violin\}}) * \\ &\quad \frac{P(List1|List1)}{P(List1)} * \frac{P(List1|Grass)}{P(List1)} * \frac{P(List1|Violin)}{P(List1)} * \\ &\quad \frac{P(Table|List1)}{P(Table)} * \frac{P(Table|Grass)}{P(Table)} * \frac{P(Table|Violin)}{P(Table)} \end{aligned} \tag{D.10}$$

Since the test lists are randomised, the history factors show no advantage for either trace. The only component that is reliably different is $\frac{P(Table|Table)}{P(Table)}$ which will always be greater than $\frac{P(Table|Grass)}{P(Table)}$. Hence, the {List1, Table, Road} trace will be chosen in preference to any of the other traces and rational analysis is able to deal with the ABABr paradigm.

In the analysis of the ABABr paradigm it was assumed that the factors that make up the event of seeing a list could be bundled together and treated in much the same way as any other cue or component of a memory trace. In this case, List1 and List2 are acting as contexts in the sense that is usual in the memory literature. Anderson and Milson (1989) make no specific comment on how these contextual factors should be handled. The above analysis is an extrapolation of the paradigm in light of the way in which these elements are dealt with in the matrix

model (Humphreys et al., 1989) and SAM (Raaijmakers & Shiffrin, 1981). An important characteristic of these factors is that they will have very low probabilities of occurrence given that it is unlikely that the same set of circumstances will recur often (particularly the same internal state). By the same argument they are unlikely to influence conditional probabilities. These characteristics will be important in the discussion of the crossed associates problem to follow.

D.4 The Crossed Associates Paradigm

The crossed associates paradigm (Spence & Lair, 1965; McCullers, 1965) involves presenting subjects with a list of paired words that consists of repairings of strongly preassociated words. For instance, the following list might be presented to subjects:

Pre-existing Associates

Bread	Butter
King	Queen

Study List

Bread	Queen
King	Butter

Two questions can be asked in this case. The first requires the subject to recall the word that was paired with a cue word in the study list. For instance, the subject might be asked which word was paired with "Bread" in the study list to which they would be expected to respond "Queen". Alternatively, the subject might be asked to recall the first word that comes to mind given "Bread". The most popular response in this case would be "Butter". Subjects are able to perform both these tasks and there is little interference from the presentation of the study list on the pre-existing associates that are reported.

In providing a rational analysis of the crossed associates paradigm one must consider both recall from the study list and recall of pre-existing associates.

Recall from the Study List

Consider the case where the subjects are required to recall from the study list. In this case, the cue set is {List, Bread}. The "List" cue is a contextual cue similar to

those used in the discussion of the ABABr paradigm. The contending memory traces are {Bread, Butter} and {List, Bread, Queen}. Expanding the need probabilities gives:

$$P(\{Bread, Butter\}|H_{\{Bread,Butter\}}\&\{List, Bread\}) = H_{\{Bread,Butter\}} * \frac{P(List|Bread)}{P(List)} * \frac{P(List|Butter)}{P(List)} * \frac{P(Bread|Bread)}{P(Bread)} * \frac{P(Bread|Butter)}{P(Bread)} \quad \text{(D.11)}$$

and

$$P(\{List, Bread, Queen\}|H_{\{List,Bread,Queen\}}\&\{List, Bread\}) = H_{\{List,Bread,Queen\}} * \frac{P(List|List)}{P(List)} * \frac{P(List|Bread)}{P(List)} * \frac{P(List|Queen)}{P(List)} * \frac{P(Bread|List)}{P(Bread)} * \frac{P(Bread|Bread)}{P(Bread)} * \frac{P(Bread|Queen)}{P(Bread)} \quad \text{(D.12)}$$

Determining which is the favoured trace in this case is not as straight forward as in the ABABr paradigm. However, since many of the factors involve conditionalizing on an irrelevant trace element they can be eliminated as they will be very near one. $\frac{P(List|Bread)}{P(List)}$, $\frac{P(List|Butter)}{P(List)}$, $\frac{P(List|Queen)}{P(List)}$, $\frac{P(Bread|List)}{P(Bread)}$ and $\frac{P(Bread|Queen)}{P(Bread)}$ can be eliminated in this way. The $\frac{P(Bread|Bread)}{P(Bread)}$ factor is common to both expressions and, hence, favours neither. The $\frac{P(Bread|Butter)}{P(Bread)}$ factor creates some uncertainty, since Bread and Butter are associates and, hence, the {Bread, Butter} trace is favoured. However, the history factors will strongly favour {List, Bread, Queen} as this has just been seen in the study list. In addition, given the arguments presented in the ABABr section, the $\frac{P(List|List)}{P(List)}$ should be well above one since $P(List|List)$ will be near one while $P(List)$ will be very small and will probably be the dominant factor. Hence, it seems reasonable to suppose that:

$$H_{\{Bread,Butter\}} * \frac{P(Bread|Butter)}{P(Bread)} < H_{\{List,Bread,Queen\}} * \frac{P(List|List)}{P(List)} \quad \text{(D.13)}$$

and that the {List, Bread, Queen} trace will be chosen as required.

Recall of Pre-existing Associates

The second question that can be asked involves cuing with just the "Bread" cue. In this case, the {Bread, Butter} memory trace should prevail. The terms involving

the List cue drop out leaving the following expressions for the need probabilities:

$$P(\{Bread, Butter\}|H_{\{Bread,Butter\}}\&\{Bread\}) = \\ H_{\{Bread,Butter\}} * \frac{P(Bread|Bread)}{P(Bread)} * \frac{P(Bread|Butter)}{P(Bread)} \tag{D.14}$$

and

$$P(\{List, Bread, Queen\}|H_{\{List,Bread,Queen\}}\&\{Bread\}) = \\ H_{\{List,Bread,Queen\}} * \frac{P(Bread|List)}{P(Bread)} * \frac{P(Bread|Bread)}{P(Bread)} * \frac{P(Bread|Queen)}{P(Bread)} \tag{D.15}$$

The terms $\frac{P(Bread|List)}{P(Bread)}$ and $\frac{P(Bread|Queen)}{P(Bread)}$ will be approximately one. The $\frac{P(Bread|Bread)}{P(Bread)}$ factor is common to both and, hence, favours neither. For the {Bread, Butter} trace to be favoured the following inequality must hold:

$$H_{\{Bread,Butter\}} * \frac{P(Bread|Butter)}{P(Bread)} > H_{\{List,Bread,Queen\}} \tag{D.16}$$

This inequality incorporates the same terms as above except that the $\frac{P(List|List)}{P(List)}$ term is missing. The $\frac{P(List|List)}{P(List)}$ term was expected to be the dominate term in the above analysis, and it seems reasonable that its absence will alter the need probabilities sufficiently to favour the {Bread, Butter} trace.

Given some assumptions about the probabilities of context cues, rational analysis is able to accommodate both the ABABr paradigm and the crossed associates paradigm. It is able to deal with joint information and, hence, passes the tests of Humphreys et al. (1989).

In order to do these tasks, three way joint information is required (Humphreys et al., 1989). In the matrix model, this information is retained by a rank three tensor. Humphreys et al. (1989) note that with minor extensions TODAM (Murdock, 1982), CHARM (Eich, 1982) and Minerva II (Hintzman, 1984) are also capable of accommodating these phenomena. In the rational model, storage of the three way bindings is a consequence of the fact that it is a trace model. As in Minerva II (Hintzman, 1984), it incorporates items into separate bundles and it is these separate bundles that maintain the necessary information.

Appendix E

Verifying the List Strength Training Set

E.1 Expressing the environmental constraints

To check that the training set constructed as specified in the method section satisfies the environmental statistics, the environmental constraints must first be re-expressed in terms of the training set conditions. There are three constraints to consider. They are the item strength constraint:

$$P(W_i \mid W_i \; is \; strong) > P(W_j \mid W_j \; is \; weak) \tag{E.1}$$

the length constraint:

$$P(W_i \mid W_i \; in \; long \; context) < P(W_j \mid W_j \; in \; short \; context) \tag{E.2}$$

and the null list strength constraint:

$$P(W_i \mid W_{k \neq i} \; is \; weak) \approx P(W_j \mid W_{k \neq i} \; is \; strong) \tag{E.3}$$

Each of the terms in the above constraints expresses the probability of a given word W_i given some condition on the history of that item. In equation 5.1 the conditions involve the strength (in this case, the number of occurrences) of the word. In equation 5.3 the condition involves the length of the preceding context (measured by the number of unique items), and in equation 5.4 the conditions involve the strength of other items in the preceding context.

As in the simulations in the HRN chapter (chapter three), the recognition task is simplified to the presentation of a list of study items followed by a single test item. The occurrence of an item is considered to be analogous to the presentation of a test item and the history is analogous to the study list.

Hence, equation E.1 is re-expressed as:

$$P(W_i \in Test \mid W_i \; is \; strong) > P(W_i \in Test \mid W_i \; is \; weak) \tag{E.4}$$

where W_i *is strong* means that W_i is presented twice in the study list and W_i *is weak* means that W_i is presented once.

Similarly, inequality E.2 can be expressed as:

$$P(W_i \in Test \mid W_i \; in \; long) < P(W_i \; \in Test \mid W_i \; in \; short) \tag{E.5}$$

Here, W_i *in long* means that W_i is drawn from the length list (i.e. eight items) and W_i *in short* means that W_i is drawn from a mixed list or from one of the pure lists.

Inequality 5.4 can be re-expressed as:

$$P(W_i \in Test \mid W_{k \neq i} \; is \; weak) \approx P(W_i \in Test \mid W_{k \neq i} \; is \; strong) \tag{E.6}$$

Again, $W_{k \neq i}$ *is weak* and $W_{k \neq i}$ *is strong* now refer to the study lists.

In the analyses which follow the length condition has not been included in the expressions concerning strength despite the fact that the items in the length condition are presented only once and hence might be considered weak. However, when the environmental statistics embodied by E.1 and E.3 were collected the length of the list in terms of unique items was held constant. Inclusion of the length condition into these analyses would introduce a systematic bias not observed in the environmental analysis. Hence the inequalities have been proven for lists with two unique items (i.e., the pure weak, mixed and pure strong lists).

E.2 The training set statistics

In presenting the proofs which follow it is useful to summarise the statistics of the training set. Table E.1 provides the probabilities of each of the list types. Table E.2 summarises the probabilities that a given word appears in a given condition and table E.3 outlines the conditional probabilities that a given word occurs as a test item given the condition.

$P(length)$	0.04
$P(pure\ weak)$	0.32
$P(mixed)$	0.32
$P(pure\ strong)$	0.32

Table E.1: Probabilities of each of the list types in the list strength training set.

$P(W_i \in length)$	$\frac{P(length) * \#\ study\ items\ chosen(1)}{vocabulary\ size(16)}$	0.02
$P(W_i \in pure\ weak)$	$\frac{P(pure\ weak) * \#\ study\ items\ chosen(2)}{vocabulary\ size(16)}$	0.04
$P(W_i \in mixed\ weak)$	$\frac{P(mixed) * \#\ study\ items\ chosen(1)}{vocabulary\ size(16)}$	0.02
$P(W_i \in mixed\ strong)$	$\frac{P(mixed) * \#\ study\ items\ chosen(1)}{vocabulary\ size(16)}$	0.02
$P(W_i \in pure\ strong)$	$\frac{P(pure\ strong) * \#\ study\ items\ chosen(2)}{vocabulary\ size(16)}$	0.04

Table E.2: Probabilities that a given word appears in a given condition in the list strength training set.

$P(W_i \in Test \mid W_i \in length)$	0.01
$P(W_i \in Test \mid W_i \in pure\ weak)$	0.075
$P(W_i \in Test \mid W_i \in mixed\ weak)$	0.15
$P(W_i \in Test \mid W_i \in mixed\ strong)$	0.3
$P(W_i \in Test \mid W_i \in pure\ strong)$	0.15

Table E.3: Probabilities that a word occurs as a test item given the condition in the list strength training set.

E.3 The item strength constraint

Required to Prove

$$P(W_i \in Test | W_i \; is \; strong) > P(W_i \in Test | W_i \; is \; weak) \tag{E.7}$$

Proof

$$\begin{aligned}
&P(W_i \in Test | W_i \; is \; strong) \\
&= P(W_i \in Test | W_i \in pure \; strong \vee W_i \in mixed \; strong) \\
&= \frac{\left(\begin{array}{l} P(W_i \in Test | W_i \in pure \; strong) P(W_i \in pure \; strong) + \\ P(W_i \in Test | W_i \in mixed \; strong) P(W_i \in mixed \; strong) \end{array} \right)^{\dagger}}{P(W_i \in pure \; strong) + P(W_i \in mixed \; strong)} \\
&= 0.2
\end{aligned}$$

$$\begin{aligned}
&P(W_i \in Test | W_i \; is \; weak) \\
&= P(W_i \in Test | W_i \in pure \; weak \vee W_i \in mixed \; weak) \\
&= \frac{\left(\begin{array}{l} P(W_i \in Test | W_i \in pure \; weak) P(W_i \in pure \; weak) + \\ P(W_i \in Test | W_i \in mixed \; weak) P(W_i \in mixed \; weak) \end{array} \right)^{\dagger}}{P(W_i \in pure \; weak) + P(W_i \in mixed \; weak))} \\
&= 0.1
\end{aligned}$$

QED

† See Appendix E.6 for proof of the conditional of ors identity.

E.4 The length constraint

Required to Prove

$$P(W_i \in Test | W_i \; in \; long) < P(W_i \in Test | W_i \; in \; short) \tag{E.8}$$

Proof

$$\begin{aligned}
&P(W_i \in Test | W_i \; in \; long) \\
&= P(W_i \in Test | W_i \in \; length) \\
&= 0.01
\end{aligned}$$

$$\begin{aligned}
&P(W_i \in Test | W_i \; is \; short) \\
&= P(W_i \in Test | W_i \in pure \; weak \vee W_i \in pure \; strong \vee \\
&\qquad\qquad W_i \in mixed \; weak \vee W_i \in mixed \; strong) \\
&= \frac{\left(\begin{array}{l} P(W_i \in Test | W_i \in pure \; weak) P(W_i \in pure \; weak) + \\ P(W_i \in Test | W_i \in pure \; strong) P(W_i \in pure \; strong) + \\ P(W_i \in Test | W_i \in mixed \; weak) P(W_i \in mixed \; weak) + \\ P(W_i \in Test | W_i \in mixed \; strong) P(W_i \in mixed \; strong) \end{array} \right)^{\dagger}}{\left(\begin{array}{l} P(W_i \in pure \; weak) + P(W_i \in pure \; strong) + \\ P(W_i \in mixed \; weak) + P(W_i \in mixed \; strong) \end{array} \right)} \\
&= 0.15
\end{aligned}$$

QED

† See Appendix E.6 for proof of the conditional of ors identity.

E.5 The null list strength constraint

Required to Prove

$$P(W_i \in Test | W_{k \neq i} \; is \; strong) \approx P(W_i \in Test | W_{k \neq i} \; is \; weak) \tag{E.9}$$

Proof

$$\begin{aligned}
&P(W_i \in Test | W_{k \neq i} \; is \; strong) \\
&= P(W_i \in Test | W_i \in pure \; strong \vee W_i \in mixed \; weak) \\
&= \frac{\left(\begin{array}{l} P(W_i \in Test | W_i \in pure \; strong) P(W_i \in pure \; strong) + \\ P(W_i \in Test | W_i \in mixed \; weak) P(W_i \in mixed \; weak) \end{array} \right)^{\dagger}}{P(W_i \in pure \; strong) + P(W_i \in mixed \; weak)} \\
&= 0.15
\end{aligned}$$

$$\begin{aligned}
&P(W_i \in Test | W_{k \neq i} \; is \; weak) \\
&= P(W_i \in Test | W_i \in pure \; weak \vee W_i \in mixed \; strong) \\
&= \frac{\left(\begin{array}{l} P(W_i \in Test | W_i \in pure \; weak) P(W_i \in pure \; weak) + \\ P(W_i \in Test | W_i \in mixed \; strong) P(W_i \in mixed \; strong) \end{array} \right)^{\dagger}}{P(W_i \in pure \; weak) + P(W_i \in mixed \; strong)} \\
&= 0.15
\end{aligned}$$

QED

† See Appendix E.6 for proof of the conditional of ors identity.

E.6 Conditional of Ors Identity

In the above the Conditional of Ors identity was used several times. What follows is the proof of this identity.

Required To Prove

$$P(A|C_1 \vee .. \vee C_n) = \frac{\sum_{i=1..n} P(A|C_i)P(C_i)}{\sum_{i=1..n} P(C_i)}$$

provided the conditions C_i are mutually exclusive.

Proof

$$P(A|C_1 \vee .. \vee C_n) = \frac{P(A\&(C_1..C_n))}{P(C_1..C_n)}$$

Distributing the & over $\vee$

$$P(A|C_1 \vee .. \vee C_n) = \frac{P((A\&C_1)..(A\&C_n))}{P(C_1..C_n)}$$

Since $P(C_i \vee C_j) = P(C_i) + P(C_j) - P(C_i\&C_j)$ and given that all of the conditions are mutually exclusive (i.e., $P(C_i\&C_j) = 0$) it follows that

$$\begin{aligned} P(A|C_1 \vee .. \vee C_n) &= \frac{\sum_{i=1..n} P(A\&C_i)}{\sum_{i=1..n} P(C_i)} \\ &= \frac{\sum_{i=1..n} P(A|C_i)P(C_i)}{\sum_{i=1..n} P(C_i)} \end{aligned}$$

QED

Glossary of Frequently Used Terms

Encodingism: A representational scheme in which the meaning of one representation is determined by composition of existing representations.

Interactivism: A representational scheme in which the meaning of a representation is determined via a goal seeking interaction between the cognitive mechanism and the environment.

Learning: The ability to create representations, decision criteria and control processes as a consequence of the interaction of a mechanism and the environment.

Learning-to-learn: The acquisition of higher order habits and skills that produce more or less permanent changes in a subjects mode of attack on a task.

Memory: The set of skills and structures that aid in the encoding, storage and retrieval of information in the human cognitive system.

Nonspecific transfer: The learning-to-learn effects that depend on the higher order skills that are independent of the relationships between successive tasks such as response and associative learning.

Null list strength effect in recognition: The finding that the strength (i.e. number of repetitions or length of study time) of other items does not affect the accuracy of recognition of a target item (Ratcliff et al., 1990; Shiffrin et al., 1990).

Specific transfer: The learning-to-learn effects that depend on similarity relationships between successive tasks.

Word frequency effect in recognition: The finding that words that are low frequency in everyday use are recognized more accurately than words of high frequency (Gorman, 1961; Schulman & Lovelace, 1970).

Index

ABAB' paradigm, 125
ABABr paradigm, 10, 50, 125, 171
ABAC paradigm, 125
ABAD paradigm, 50
accommodation, 2
artificial intelligence, xxi, 2
assimilation, 2
associative learning, 17, 126
associative practice, 16

backpropagation, 7, 12, 18, 55, 80, 130
backpropagation through time, 62
Bayesian odds formula, 167
BPTT, 62
Burrel model, 168

canonical discriminants analysis, 85
capacity, 59
catastrophic interference, 59, 78, 92
CDA, 85
CHARM, xviii, 7, 57, 58, 65, 176
closed-loop system, 62
cognitive psychology, xxi
cognitive science, xxi, 153
computational constraints, xxi
concept learning, 12, 155
concreteness, 27
conditional of ors identity, 183
connectionism, 3
connectionist mailing list, 33
context, 69, 172
context factor, 20, 167
contextual frequency, 110
control processes, xxi, 6, 58, 91, 122, 124
convolution, 8, 57
correlation, 57
cross domain influences, 130
cross paradigm generality, 17, 127
crossed associates paradigm, 10, 50, 174
cued recall, 39, 50, 129
culture, 18

d', 74
decision criteria, xxi, 6, 58, 91, 122, 124
development, 15
differentiation, 39
domain specific predispositions, 15
dot product, 8, 57

elaboration, 15
electronic mail, 41
encoding process, 68, 131
encoding specificity, 131
encodingism, 156, 185
encodings, 57, 122

environment, 18
episodic recognition, 32
evolution, 18

familiarity, 8
feedforward network, 68
fixation of belief, 155
forced choice recognition, 27, 36
free recall, 16, 39, 129
functor words, 33, 159

gaussian, 10
general repetition rate, 45, 47, 165
generalization, 60, 80, 103, 106
global matching models, 7

HCA, 85
Hebbian memory, 91
hebbian recurrent network, xxi, 56, 61, 97, 109, 114, 122, 123
hidden units, 68, 83, 91
hierarchical cluster analysis, 85
history factor, 20, 168
HRN, xxi, 56, 61, 97, 109, 114, 122, 123
HRN algorithm, 66
HRN architecture, 65

images, 9
imaging, 15
information processing paradigm, 1, 153
interactive model, 56, 108, 109
interactivism, 22, 94, 121, 123, 126, 156, 185
interference, 4, 13, 59, 91
item strength effect, 110, 113, 122, 177

Jordan network, 62

learning, 5, 185
learning rate, 96, 113, 128
learning theory, 1
learning-to-learn, 3, 15, 61, 124, 185
lexical decision, 27, 30, 37
list length, 75
list length effect, 111, 113, 122, 177
list strength training set, 177

matrix model, xviii, 7, 10, 40, 58, 65, 122, 126, 173, 176
mean word density, 34
meaning, 157
meaningfulness, 16
medium term store, 72
memory, 61, 185
memory decay rate, 79, 96, 113, 128
memory development, 123
memory skills, 15
memory trace, 11, 167, 171
metamemory, 58
Minerva II, xviii, 7, 11, 58, 65, 126, 176
mirror effect, 28
mixed-pure paradigm, 170
MTS, 72

naming, 15
nativism, 154
need probability, 20
negative evidence, 133

New York Times, 41
nonspecific transfer, 16, 185
null list strength effect, 25, 38, 109, 112, 113, 121, 122, 128, 133, 169, 177, 185
number of associations, 27
number of effective parameters, 103
number of hidden units, 128

orthogonality, 59, 78
orthographic distinction, 27

paired associate learning, 16
parental speech, 41
parsimony, 128
PCA, 83
phrasal concepts, 155
physical symbol system, 153
pre-experimental recency, 27
principal components analysis, 83
principle centered approach, 128, 131
proactive interference, 16, 27, 37, 107

rapid binding, 61, 92
ratio of ratios, 116, 169
rational analysis, 19, 126, 167
reaction time, 129
reading, 32
real time recurrent learning, 62
recall, 8, 129
recognition, 7, 8, 26, 32, 37, 39, 70, 129, 133
recurrence probability, 47, 165
recurrent networks, 59
rehearsal, 15
representation, xxi, 6, 57, 91, 124, 157
response integration, 16
response learning, 16, 126
retrieval process, 68, 131
retrieval structure, 9
retroactive interference, 59
robot arm, 18
RTRL, 62

SAM, xviii, 7, 8, 28, 39, 58, 64, 126, 174
schooling, 133
serial position, 11, 21
serial recall, 16, 129
signal detection theory, 8, 58, 74
similarity, 127
similarity landscape, 124
simple recurrent network, 62, 122, 127
specific transfer, 16, 185
SRN, 62, 122, 127
Star Trek, 50
storage, 4, 68
structural distinction, 27
study list, 7, 70
symbol grounding problem, 154
symbol system, 154
symmetric sigmoid, 78, 97

target repetition rate, 45, 47, 165
task centered approach, 128
TDNN, 62
tensor memory, 176
tensor product, 60

test list, 7
time complexity, 82, 91
time delay neural network, 62
TODAM, xviii, 7, 41, 58, 65, 176
training set, 96
trigrams, 16

University of Minnesota Daily, 46, 159

verbal learning, 3, 15, 64
vividness, 16
vocabulary, 60, 82, 92

warm up, 16
weight decay rate, 104, 106, 113, 128
word density, 34
word frequency, 47, 165
word frequency effect, 25, 26, 93, 121, 128, 133, 167, 186