ACQUIRING AND GROUNDING A LEXICON WITH ART: TOWARDS ROBOTIC SYSTEMS THAT UNDERSTAND LANGUAGE.

by

Nathan James Chandler

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

PhD

The Nottingham Trent University

2000

Abstract

This thesis reports on the development of a method to enable artificial systems, specifically robots, to autonomously acquire a diverse lexicon of natural language terms. In this case English words which can be used to facilitate communication between themselves and their human operators. For reasons that will become clear, this task is referred to as 'lexical acquisition and grounding.'

It is argued herein that in order for this communication to succeed, the artificial systems must gain an intrinsic understanding of the meaning of these words. Moreover, it is argued that such an intrinsic understanding is ultimately founded upon the systems ability to experience, via sensory and motor systems, its surrounding environment. Put in simple terms, this is similar to arguing that a blind man could never truly understand an English word such as 'green.' Support for these arguments is provided by way of a comprehensive analysis and interpretation of a variety of recent theories (e.g. Harnad's Symbol Grounding Theory) from Cognitive Science, Psychology, and Philosophy which reject a number of 'traditional' approaches to defining the meanings of words on various grounds.

A review of related work deals with a number of practical models that have been developed from similar theoretical foundations. Based upon an assessment of the strengths and weaknesses of each approach an alternative approach to the lexical acquisition and grounding task is suggested. This approach applies an existing theory of computation – Adaptive Resonance Theory (ART) – to this task. This new application of ART is examined empirically and then ART models are adapted and extended in relation to various task specific requirements.

The aim and objectives of the work described in this thesis have been achieved, in that a model has been developed that is able to autonomously acquire English words in an incremental and continual manner and demonstrates that it has an understanding of these words. Moreover, it has been shown that this model can concurrently learn meanings with respect to more than one sensory domain and that the model can operate in a manner that is fast enough to warrant its practical implementation. With support from empirical investigation and theoretical analysis it has been shown that, on several counts, the application of an ART-based framework to the lexical acquisition and grounding task provides significant advance on systems that have been developed in the past to address this problem.

Acknowledgements

And finally, although it appears at the beginning, here is that part of the thesis that is always left to the end. Once the writing is *eventually* over one can sit down, relax and contemplate all those people that have been so supportive (i.e. moaned at me to finish) over these past years, and then try to devise the appropriate words to convey the thanks and appreciation that I owe to them all. A page is surely never enough!

First and foremost a few lines must go out to they who bought me into this world, and who have supported me through thick and thin ever since. I couldn't wish for a better mum and dad. I love you both. Always have, always will. That sentiment extends to my brother and sister, and to the wider family – I'll try to forget less birthdays etc. now that I've got less on my mind!

My three supervisors, without whom this work would never have reached fruition: Dr. Vellupillai Balendran (Bala) whose help in all things personal and professional has always been so invaluable. Our laser lab debates that have left me feeling horse on more than one occasion will always be remembered, Dr. Lindsay Evett whose advice, guidance, editing and correcting have all been vital components of my research success, and last but by no means least Dr. Kandiah Sivayoganathan (Siva) who provided me with an opportunity open to so few, this thesis is the result. I owe it to you.

To friends and colleagues far and wide, each as supportive as the next, love and respect goes out to you all (even those of you whom have often jibed about my many years in academia!!).

On an academic note I would also like to recognize and acknowledge the ever seminal work of Gail Carpenter, Stephen Grossberg, and their various colleagues at Boston University. The politics of academia seem to have dealt a poor hand to their work. However, I am sure that the time will come when it receives the wide recognition that I believe it deserves.

And saving the best till last as one always does I have only one thing to say to the person most important to me

i love u Rach.

Contents

List of Abbreviations

A glossary of common terms, symbols and abbreviations used throughout the main text of the thesis is provided as a point of reference below

ART	Adaptive Resonance Theory
ARTMAP	Class of Supervised Learning ART Architectures
CM	Connectionist Model
ENLP	Embodied Natural Language Processing
ES	Embodied Semantics
FA	Fuzzy ARTMAP
FACC	Fuzzy ARTMAP (with Complement Coding)
GA	Gaussian ARTMAP
HSV	Hue, Saturation & Variance (Colour Model)
LAG	Lexical Acquisition & Grounding
MLP	Multi-Layer Perceptron
MM	Muti-Modal
RGB	Red, Green & Blue (Colour Model)
SN	Semantic Network
TOM	Triange of Meaning
WANN	Weightless Artificial Neural Network

Chapter 1

Introduction

1.1 Background

The advent of general purpose computing technology has revolutionised the way that modern society works. In the field of engineering its impact has been profound, having brought about advances in diverse areas such as design (computer aided design), production scheduling, robotic control, and inventory management to name but a few.

In recent years a great deal of effort has been focused toward making the powerful artificial systems that employ such technology easier to use. The interface between human and computer, man and machine, has become a topic of major research interest. One particular vision that continues to motivate a great deal of the contemporary work in this area is that of natural language communication; the realisation of artificial systems that can understand and utilize the language that humans use to communicate, e.g. spoken and written English.

The benefits to be gained from building artificial systems that can 'speak our language' are innumerable and far reaching. For one, such technology would be inherently more accessible. Currently, in a typical engineering environment, the operation of most computerised production systems, e.g. robotic cells, machine tools etc., entails first acquiring an understanding of the formal artificial language of control that the particular system exploits. Natural language interaction, e.g. via spoken English, would eliminate this prerequisite and in doing so reduce the burden of training in terms of both time and cost.

The academic study of human-computer natural language communication has become established within a field known as Natural Language Processing (NLP) which is itself a sub-field of Artificial Intelligence (AI) [1]. The NLP research field is far from autonomous; Theoretical Linguistics, Cognitive Psychology, and Philosophy are but a few, amongst the diverse body of disciplines that motivate and influence NLP practice.

Despite decades of active NLP research, e.g. see Grishman (1986), the scientists and engineers who work in the field are still a long way from constructing an artificial system that can truly be said to understand even a modest proportion of natural language. In fact, the current state-of-the-art NLP systems still demonstrate a linguistic aptitude greatly inferior to that of an average pre-school child. Although a 'full solution' has not been forthcoming, significant advances have been made (Bates & Weischedel 1993, Allen 1995, Jurafsky & Martin 2000). These have, by and large been confined within specific sub-areas of NLP proper, these sub-areas being :

- *Phonetic Analysis* : Considers how the basic sounds (phonemes) of spoken language link to produce words e.g. the word 'create' is constructed from the English phonemes *k, r, ee, ay, t.*
- *Morphological Analysis* : Considers the various constituents, or morphemes that make up a given word, e.g. the adjective 'incoming' is formed from 'in' (a prefix), 'come' (the root), and 'ing' (a suffix).
- *Syntactic Analysis* : Considers the set of structural relationships (a grammar) that hold between the various syntactic classes of words (nouns, verbs, adjectives etc.) in a given language.
- *Semantic Analysis* : Considers the interpretation (meaning) of linguistic expressions (symbols) from the lexical level (words) through to the sentential level.

[1] Natural language ability is often cast as the fundamental defining feature of general intelligence. This view is evident in the *Turing Test* which was proposed, by the famous British scientist Alan Turing, as a measure by which to judge intelligence. A human judge is required to converse with another agent via a teletype system. Based upon the written responses the judge is required to determine if the agent is human or artificial. If an artificial agent is judged to be human then in Turing's eyes the agent could be viewed as being intelligent.

- *Pragmatic Analysis* : Considers the intentional, as opposed to literal, meaning of a linguistic expression, taking into account the surrounding communicative situation (discourse) and knowledge of the world.

In the area of phonology, a great deal of academic research has now developed to the point of commercial fruition, e.g. phonological findings have found application in the latest Automatic Speech Recognition (ASR) systems. The same is also true of morphological research; a number of practical large-scale analyzers are now commercially available (a review of applied morphological analysis can be found in (Karlsson & Karttunen 1996)). By contrast, progress in syntactic, semantic, and pragmatic analysis has been less forthcoming and many problems still remain.

1.2 Domain of Investigation

The work described in this thesis addresses certain issues that are of fundamental importance within the domain of semantics. Specifically, it considers how artificial systems may acquire an intrinsically meaningful set of basic linguistic symbols. This task — *lexical acquisition and grounding (LAG)* — is now described in greater depth.

1.2.1 Basic Linguistic Symbols

Basic linguistic symbols are the minimal meaningful units of quantity within a language, e.g. phrases, words, morphemes. Throughout the discussions detailed in this thesis, only words are considered as candidates. Based upon this consideration, a set of these basic symbols can be viewed alternatively as a lexicon. In the original investigations that follow in the latter sections of the thesis, the lexicon that is acquired is restrained to a collection of English nouns. The rationale for this is twofold :

1. The work presented here is put forth as a step toward natural language interaction between humans and robotic systems. Typically, in engineering environments such systems interact with solid objects in a three dimensional domain (e.g. robotic assembly tasks (Howarth 1998)). For language to be of any use the lexicon employed

must suffice to describe such a domain. In English, the words that are regularly used to refer to such 3D objects fall within the syntactic category of symbols labelled as nouns.

2. From an alternative perspective, members of the syntactic category of nouns — such as those used to refer to 'concrete' objects (Sales 1996) — are generally the first lexical symbols that are acquired in human linguistic development (Clark 1973, McShane 1979).

The basic linguistic symbols are of extreme importance from a foundational point of view, since it is the meanings of these basic symbols that combine to give meaning to the more complex composite linguistic symbols, e.g. sentences. Without some understanding of the words 'block', 'pick', and 'red', the possibility that one might understand the sentence, 'pick up the red block,' seems remote (cf. the principle of compositionality [see section 2.1.1]).

1.2.2 Intrinsic Meaning

Written words such as those that combine together to make the text on this page have no meaning in and of themselves, they are simply arbitrary patterns and squiggles. For example, the word 'dog' does not embody any feature or property of the distal objects to which it refers, i.e. it has no smell, feel, sound, shape etc. that bears any relation to "man's best friend" [2]. Despite this fact, when processed by a competent English language user, the text *is* meaningful because each pattern or squiggle has a consistent interpretation, i.e. the user can associate a word with a concept (the users intrinsic semantic knowledge).

The general approach to building the semantic component of artificial NLP systems has traditionally relied heavily upon designer introspection. The common view is that one can endow an artificial system with semantic knowledge — in the form of explicit symbolic representations — devised on the basis of one's own conscious understanding of the way the world is, i.e. the designer strives to formalise (symbolise) his, or her own knowledge. Figure 1.1 provides an example of how the meaning of, or the concept referred to by, the

[2] It is exactly this property (arbitrariness) that gives words their symbolic status — words are quintessential symbols (cf. (Pinker 1994, Dorffner 1995)).

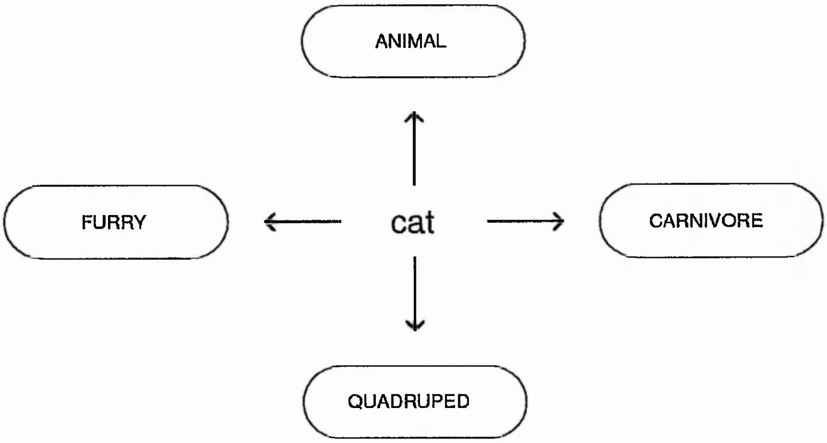

Figure 1.1: Semantic knowledge: a (pseudo) symbolic representation of the meaning of 'cat.'

word 'cat' might be explicitly represented[3] in terms of other symbols.

In recent years, this type of approach has come under increasing scrutiny. Arguments abound[4] that maintain a full account of meaning cannot be captured by way of such symbol to symbol relations — they all point to an underlying paradox which is integral to the above methodology: *meaning can be defined in terms of meaningless entities.* A brief explanation of this paradox can be put as follows. When processed by humans, explicit representations of the type described above fulfil their function because, and only if, the definiendum (the symbols to which the unknown symbol is related) are meaningful entities[5], e.g. a child that has never encountered the symbol 'cat' can learn about its reference based upon his/her implicit prior knowledge of the symbols 'ANIMAL', 'FURRY', etc. In an artificial NLP system one cannot assume this to be the case. In fact the definiendum are just more arbitrary squiggles that are in need of an interpretation themselves, i.e. 'ANIMAL', 'FURRY', etc., are as meaningless as 'cat' to the artificial system. Based upon the above approach, an interpretation of these definiendum can only be established by referring them to further symbolic definiendum, which will again be in need of interpretation. The point is, as

[3] The systems of logic, semantic networks etc. that have traditionally provided the infrastructure for such knowledge representation are described in greater detail in chapter 2.

[4] The most historically consequential of these, and other arguments that position themselves against a symbolic account of meaning representation, are given the full consideration that they deserve later in the thesis.

[5] The functionality of a dictionary is dependent upon this assumption.

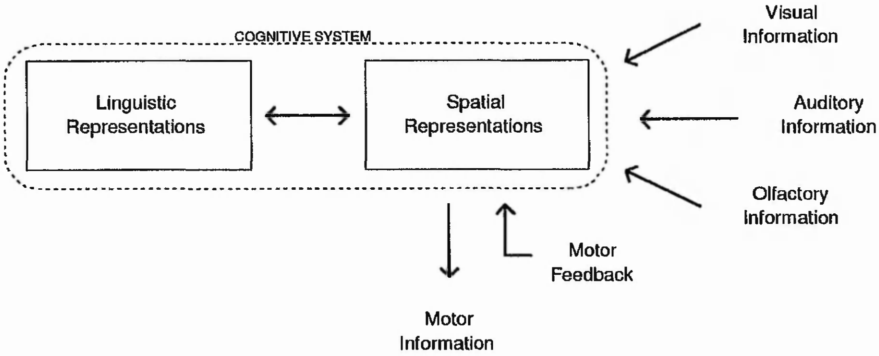

Figure 1.2: Linguistic representations (e.g. lexical symbols) have intrinsic meaning because they are cognitively submersed (figure adapted from (Landau & Jackendoff 1993)).

Dennett so eloquently puts it, that "a computer whose only input and output was verbal would always be blind to the meaning of what was written" (Dennett 1969, p.182).

The work described in this thesis takes as its lead a quite different semantic methodology which is founded on the premise that the symbols which represent categories to humans (like nouns in natural languages), are intrinsically meaningful because they are connected to the world (a hubbub of electromagnetic stimuli) in a causal and non-arbitrary way, via perception — the linguistic symbols are *grounded* (Harnad 1990). From this alternative perspective, the word 'cat' is rendered meaningful because it is hooked to a mental state (an internal concept) that analogically embodies — based on the transduction and quantisation of sensory signals — some structure of the distal object, or category of objects to which the word refers, e.g. the sort of shapes that cats are, the sounds they make, the smell they have etc. Hence, on the basis of the new approach, linguistic capacity is intertwined with, and dependent upon, a wider range of cognitive functionality, i.e. input and output transcends the level of verbal description (see figure 1.2).

1.2.3 Acquisition

The majority of work that has, and continues to be conducted within the realm of NLP, and to a large extent within AI as a whole, can be thought of as designer-centric (DC), i.e. artificial system designers endeavor to encode their own knowledge within artificial models of the world using explicit rules and representations[6]. Although the DC methodology enjoys predominance within NLP practice it is certainly not above criticism. There are two somewhat overlapping critiques that are commonly voiced :

1. That DC imparted knowledge is 'brittle.' In symbolizing his or her own knowledge, the designer can often overlook a potential dependency[7]. The result of this can be the creation of fragile systems which are not robust in real environments; systems that can fail catastrophically and unexpectedly, as Harnad illustrates in the following example :

   > A "scene-understanding" program will blithely describe the goings-on in a visual scene and answer questions demonstrating its comprehension (who did what, where, why?) and then suddenly reveal that it does not "know" that hanging up the phone and leaving the room does not make the phone disappear, or something like that. (It is important to note that these are not the kinds of lapses and gaps in knowledge that people are prone to; rather, they are such howlers as to cast serious doubt on whether the system has anything like "knowledge" at all.) (Harnad 1990, p.339)

2. The manual construction and implementation of rules and representations can be tedious and error prone. As new knowledge is added — to overcome previously unforeseen knowledge gaps — the designer must check that it is not in conflict with any of the knowledge that already resides.

An alternative proposal, and one that is endorsed here, is for an artificial system to be crafted in such a manner as to enable it to automatically obtain a large proportion of the knowledge that it will require to communicate by way of natural language. This will necessitate, amongst other things, the acquisition of (1) lexical symbols, and (2) their meanings. The notion is that if the system can develop and adapt its knowledge, or understanding, in a

[6] The DC approach is also often labelled as 'representationalist.' The best contemporary large scale example of the DC approach can be found in the CYC (Lenat 1990, Lenat 1997) project which stresses natural language understanding as one amongst its numerous goals.

[7] The *Frame Problem*, which is well documented within AI (McDermott 1976, Pylyshyn 1987), is the problem of failing to clarify and formalize such dependencies.

bottom-up fashion, progressing from the rudimentary to the complex, then it is more likely to succeed where DC systems have failed, i.e. it should be robust in real environments.

The above proposal fits with, and is in part based upon, the assumption that it is the experience of, and interaction with, one's environment that motivates communication and consequently natural language. Hence, in systems that apply the above methodology, a lexicon develops as required. As a system encounters, and acquires an understanding of novel phenomena arising within its environment, it is motivated to simultaneously acquire a novel lexical symbol that can be employed to refer to these phenomena. This incremental route of learning should remove a huge burden from the system designer, viz. the need to speculate about what words and knowledge might be required[8].

It is now widely accepted that natural language skills piggy-back on a diverse and vast foundation of general knowledge. The scope of this knowledge effectively rules out the possibility that one might create an artificial system that is able to competently utilise natural language using the DC approach, and lends support to Dennett's premise that "the only practical way of doing it [creating such a system] is one version or another of Mother Nature's way — years of embodied learning." (Dennett 1997, p.359)

1.2.4 Embodied Artificial Systems (EAS)

As stated briefly above, the core motivation for the work presented in this thesis arises from the (still distant) image of advanced natural language interaction between humans and robotic-type systems. It is these robotic-type systems that can be thought of as embodied in the following sense: they are systems that can, by way of their sensorimotor peripheries, experience and/or interact in, a physical world.

Although, a robotic-type domain was envisaged from the outset, to provide a framework from which to investigate issues in human-machine interaction, its importance *vis-à-vis* natural language semantics, and understanding (from a general perspective) was not appreciated. What has been touched upon above, and shall be further clarified in the discussions that follow, is the fact that only embodied systems have any chance of truly understanding

[8] It should be noted, however, that such learning will not necessarily overcome the problem of conflicting knowledge as stated in critique 2.

natural language — given the premise that meaning is inextricably linked with, and derived from, real experience — and then only if the structure of this embodiment is apt.

This human-EAS research direction provides a stark alternative to that which has influenced the majority of NLP work, viz. that language can be thought of as a predetermined system of rules and representations that operates in a closed autonomous manner, independent of any other cognitive functionality, e.g. visual processing. Though this view is perhaps strange, when one considers that before such work was initiated one of the forefathers of modern AI, Alan Turing, seemed to suggest that as a prerequisite to natural language use, machines would require the ability to learn and perceive their surrounding environment:

> "We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like playing chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc." (Turing 1950)

It has taken many years, but finally ideas that fit well with Turing's vision are being put into NLP practice. In the last decade this alternative methodology has thrived. Supported on theoretical grounds by the arguments detailed in the following chapter it has gained greater momentum, and has 'borne fruit' in the form of several applied NLP models (e.g. (Dorffner 1992, Plunkett et al. 1992, Nenov & Dyer 1994, Sales 1996)). The work that is presented in this thesis is an attempt to embrace and extend the current state of the art within this emerging research field.

1.3 Aim and Objectives of the Study

The aim of the work detailed in this thesis is to produce a system that is able to autonomously learn English Words.

The objectives were:-

1. To produce a system in which such acquisition (or learning) can be incremental and continual.
2. To produce a system in which the acquired words have meaning that is intrinsic to the system. This is in the sense that an understanding of words is grounded in relation to the system's own sensory experience.
3. To produce a system that can demonstrate that its intrinsic understanding of words is similar to that of humans – thus facilitating communication.
4. To produce a system that is able to address the aims stated above in a practical context where responsiveness, accuracy and the efficient use of resources are all important considerations.

These are now considered in greater detail by outlining problems that need to be addressed and steps that need to be taken:

1. **Autonomous acquisition of English Words**: In this sense a model must be sought that is able to learn for itself without the need for constant intervention, tweaking and/or re-programming. Consideration is required as to how words should be presented, or fed to the system, e.g. should they be provided via a voice, textual, etc. interface? Will some form of encoding be required?
2. **Incremental and Continual Learning**: If the system developed is to be of real use in a future robotic system it is considered here that the breadth of language that can be used should not be fixed and finite for all time. As is the case with humans and language, a vocabulary should by developed and adaptable throughout life. Thus, careful consideration must be taken to ensure the learning methodologies employed in the system to be developed conform to these requirements.

3. **Attain intrinsic meaning through sensory groundings**: When a parent recites a word and points to an object in order to teach their child meaning, one imagines that the child is grounding its recognition of this word with respect to some aspects of the scene it is visually processing. In terms of developing an artificial model akin to this process one must consider a number of questions:

    - What sort of sensory experience is required?
    - How can words be mapped to sensory experience?
    - Do words map to some higher level mental concepts that derive from sensory experience?
    - How can input from multiple sensory sources be combined, e.g. to facilitate grounding with respect to shapes, colours, sounds, tastes etc.

4. **Demonstrate understanding similar to humans**: Intrinsic meaning, based upon sensory experience, is very much a subjective quantity. However, for communication to succeed one must have inter-subjective meaning. For example, there is little hope that person A will correctly respond to person B's request "point to the red one" unless both subjects share some consensual understanding of the word 'red'. A number of quantifiable tests will be required to ensure that the model developed is acquiring meanings that are *useful* in this sense.

5. **Consideration of practical applicability**: The aim is to create a model that can be realistically deployed in a practical context. In this sense it is pointless if a model *can* demonstrate understanding and correctly respond to a natural language command, but to do so takes many minutes or hours. Moreover, one expects to be able to deploy a solution on affordable hardware without seriously impacting upon performance. In particular, during the development cycle special attention will be afforded to:

    - How long it takes a model to learn words and meanings ?
    - How long it takes to recall such learning (and demonstrate understanding)?
    - How much computing resource is required for such learning and recall ?

1.4 Thesis Structure

The remainder of this thesis will cover the issues that have been touched upon above in greater detail. The various chapters of the thesis are set out as follows :

Chapter 2 is the first of two chapters that survey the academic literature connected to and surrounding the field of inquiry set out above. The chapter addresses three main areas; First, it reviews the domain of semantic analysis, providing an overview of the main classes of traditional semantic theory. Next, it considers a number of problems related to these traditional views. Finally, it introduces a new idea that addresses these problems and provides a novel theoretical framework for their interpretation.

Chapter 3 continues the consideration of related work on two lines. Initially, the text discusses in greater detail how connectionist models (Artificial Neural Networks), provide a suitable substrate for the theoretical ideas introduced in the final section of chapter 2. The bulk of the chapter then introduces and critiques a variety of applied systems that have been developed in a similar vain to the novel work presented in this thesis.

Chapter 4 proposes Adaptive Resonance Theory (ART) as an alternative and superior medium by which to achieve the objective of acquiring an intrinsically meaningful lexicon. ART encompasses a family of non-symbolic connectionist models that share certain fundamental characteristics. These characteristics are described and their applicability to the task at hand are argued.

Chapter 5 is the first of three chapters that highlight the original empirical investigations that have been undertaken. In this chapter the task focuses upon the acquisition of colour terms; a lexicon of English nouns that stand to represent colours, e.g. ‘red’. The goals that drive this research are threefold : (1) to determine the importance of embodiment, (2) to assess the effect of learning, i.e. supervised vs. unsupervised, and (3) to demonstrate and justify the application of Adaptive Resonance models.

Chapter 6 takes the work of Sales (1996), as a basis for comparative study. ART based models are applied to Sales’ original data. The results are presented and discussed in detail.

Chapter 7 investigates the acquisition of an extended lexicon which encompasses nouns that relate to either colour (chapter 5), or form (chapter 6). The first part of the chapter highlights a number of problems that rule out the application of standard ART models with respect to this task. Based on an appraisal of these problems, an adapted and improved architecture and associated learning algorithm — based upon ART principles — is put forward. A number of empirical results are presented and discussed.

Finally, Chapter 9 reviews the thesis, considering its contribution to the encompassing research field. Questions which have arisen from the work are then discussed, and a number of possible directions for further work are highlighted.

Chapter 2

Semantics and Meaning

> **Semantics** - The branch of linguistics that deals with meaning; (the study or analysis of) the relationships between linguistic symbols and their meanings.
> – *Concise Oxford English Dictionary*

As the definition above suggests semantics is devoted to the investigation of meaning, particularly how language expressions convey meaning. Consequently, based on the presumption that *understanding* concerns some transfer of *meaning*, one can assume that a semantic system is a vital and necessary component of any language understanding system: natural or artificial.

The purpose of this chapter is to provide an insight into the broad theoretical foundations upon which the empirical research described in this dissertation rests. These foundations take the form of a contemporary lexical semantic methodology that embodies various new ideas and principles. This methodology endeavors to provide answers to questions that arise from various theoretical and empirical analyses of what shall be termed herein as 'traditional' approaches to semantics and meaning.

In the ensuing section, an overview of semantics and meaning is presented which is followed by a brief introduction of the various semantic theories that fall under the 'traditional' label mentioned above. The penultimate section provides a discussion of various problems that in some way effect these traditional theories. Finally, the closing section of the chapter details various theoretical ideas that have been proposed in answer to these outstanding problems, and presents a novel graphical framework that intuitively conveys these ideas.

2.1 An Overview of Semantics and Meaning

To provide an outline and structure for an overview of semantics and meaning one can do little better than to quote what Gazdar & Mellish (1989) regard as the four fundamental questions any complete semantic theory should strive to answer :

1. What are the appropriate sub phrases to consider when we want to obtain the meaning of a phrase?
2. How does the meaning of a particular phrase depend on that of its sub-phrases?
3. What are the meanings of the minimal units (the phrases, words or morphemes), that do not themselves subdivide into smaller meaningful units?
4. What kinds of things should meanings be anyway? Are they symbols in the machine? Are they things in the world? Are they some kind of relationship between the two?

Although a complete semantic theory is ultimately sought, in practice due to the complexity of the issues involved in tackling the above questions, the practitioners of semantics, including computational linguists and AI researchers, have divided their domain of interest into specific sub-fields. A brief description of these follow.

2.1.1 Structural Semantics

The first two questions Gazdar and Mellish pose fall within the realm of what has come to be known as structural (Partee 1981), or compositional semantics. This component of a semantic theory deals with the effects a formal set of rules (i.e. a syntax) has in relation to meaning. The foundations derive in part from the Fregean Principle of Compositionality (POC) which asserts that the meaning of a sentence is a systematic function of the meaning of its components. The POC has been widely adopted amongst semantic theories. However, there are notable exceptions, for example, Procedural Semantics (Woods 1968).

2.1.2 Lexical Semantics

The third question put forward by Gazdar and Mellish relates to another sub-domain of semantic theory, namely lexical semantics. The task within this sub-field of semantics is to determine the meanings of the words (lexical components) within a language. For example, one would hope to equate the word 'orange' with a roundish, typically orange and juice laden fruit that is borne by certain types of tree, with a perceptual experience which is correlated to a certain spectral band of wavelength of reflected light photons, and also with a technology and communications company operating in the UK cellular phone market.

Following from chapter 1 it should be clear that the investigations detailed within this dissertation are primarily concentrated within the domain of lexical semantics[1]. However, the work presented herein differs from the traditional approaches within this field of study because of the way it accounts for Gazdar and Mellish's fourth question — What kinds of things should meanings be anyway? Traditionally, semanticists have divided themselves into three camps regarding their response to this question :

Meanings are Things in the World (MTW)

The terms that are used throughout the relevant literature (linguistics, philosophy, cognitive science etc.) to describe those theories that adhere to some form of notion whereby meaning resides in the world are abundant, e.g. denotational, referential, realist, and objectivist. In these types of theory the meaning (reference, denotation, extension) of a basic lexical symbol is equated with an entity, event, or state (or a category of entities, events, or states) that is assumed to exist objectively in the world. Thus, as figure 2.1 shows, the symbols 'Felix' and 'Tom' are meaningful because they stand to represent two individual entities that exist, and the symbol 'cat' is meaningful because it represents a category (or set) of entities that exist. These ideas are clarified, from an implementational perspective, in section 2.2.1 below.

[1] It should be noted that the term lexical semantics is slightly ambiguous in that it is also used in some circles to describe what is more appropriately identified as word-sense disambiguation. Confusion can arise here because this process is actually structural in nature. Given that a lexical symbol has a number of alternative meanings (or senses), the task is to assess which of these fits with the context of the surrounding content. For example, in the sentence 'the boy peeled the orange in delight,' the object of the disambiguation exercise is to attain the fact that the word 'orange' refers to a piece of fruit.

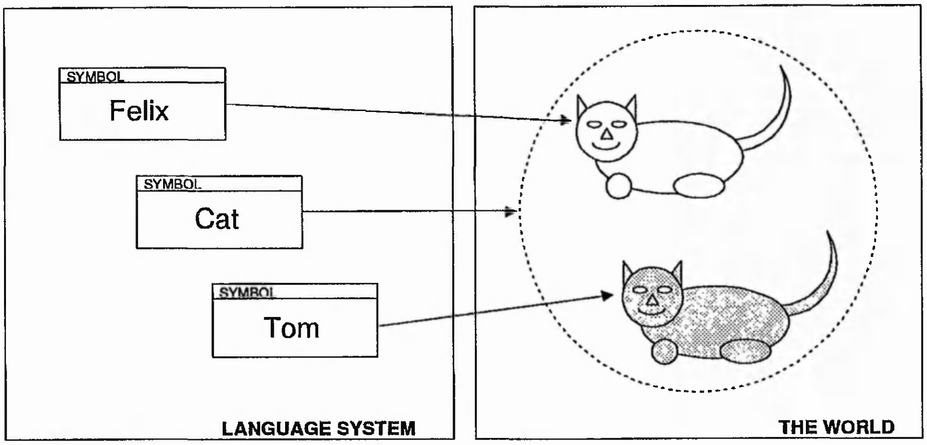

Figure 2.1: An objectivist (MTW) account of meaning.

Meanings are Symbols in the Machine (MSM)

Another view that is prevalent in traditional semantics is that the meaning (sense; connotation; intension) of a lexical symbol is equated with a concept that is activated in the mind of the understander, i.e. a mental state. These mental states are thought to be embodied by way of the contractual relations one symbol has with other symbols, e.g. the meaning of the symbol 'cat' might be defined through its relations with the symbols 'animal', 'furry' etc. (see figure 1.1 in chapter 1).

One issue that remains unresolved by traditionalists is whether the set of symbols to which a symbol is related should be open (relations can be defined between any of the lexical symbols), or closed (there is some set of primitive symbols to which all others are related). These alternatives will become clearer in light of an overview of traditional semantic theories which follows.

Meanings are Symbols and Things (MST)

One final point of view that Gazdar and Mellish highlight as a question above is that which integrates MTW and MSM. According to this view the contractual relations a lexical symbol has with other symbols, specify the necessary and sufficient conditions for reference; a lexical symbol refers to some entity in accordance with a mediating concept. More shall be offered in relation to this view later.

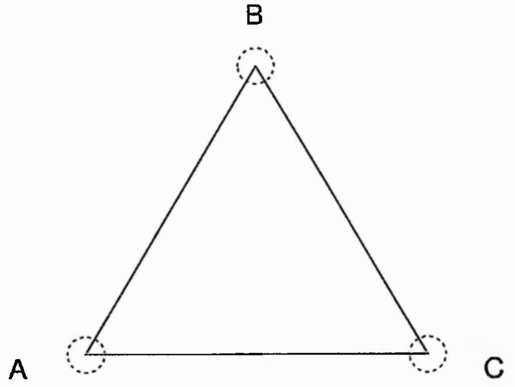

Figure 2.2: Ogden and Richards triangle of meaning

The Triangle of Meaning

The essence of the various positions held *vis-à-vis* the question of meaning are captured succinctly by Ogden & Richards (1923) triangle of meaning (TOM) — a diagrammatic tool, depicted in figure 2.2, that has been employed extensively in semantic debate. The TOM is introduced here because it serves well to clarify important issues that arise later in the thesis. The vertices of the triangle represent a hierarchy of semantic levels, following Lyons (1977), these vertices are denoted using the letters A, B, and C. In the context of the preceding discussions, A can be substituted by *symbol*, B by *concept*, and C by *entity*, and the relations AC, AB, and ABC characterise the pure MTW, pure MSM, and integrated MST perspectives respectively.

2.2 Semantic Theories

With a brief overview of the terrain complete, discussions now turn to the 'traditional' semantic theories which have been influential in crafting aspects of the various theoretical models of meaning ventured above, and which have, and continue to provide the frameworks that support the implementation of these models.

2.2.1 Models, Sets, and Meaning Postulates

Model-theory (MT), or model-theoretic semantics, utilizes formal mathematical models (which are thought to characterize the *real* world) in order to create a correspondence between linguistic expressions and the entities to which they refer. A model in this context is an abstract data structure that consists of :

1. U , a set of semantic objects called individuals that represent all entities within a domain of discourse, e.g.

$$U \quad = \quad \{\, 0,\, 1,\, \mathit{Barry},\, \mathit{Nathan},\, \mathit{Rachael}\,\}$$

2. V , a function assigning values to constants x. For individual constants, $V(x)$ is an individual (an entity within U), or set of individuals. If x is an n-place predicate, $V(x)$ is a set of n-tuples of individuals, e.g.

$$\begin{array}{lcl@{\qquad}lcl} V(\text{a}) & = & \mathit{Barry} & V(\text{male}) & = & \{\,\mathit{Barry}; \mathit{Nathan}\,\} \\ V(\text{b}) & = & \mathit{Nathan} & V(\text{female}) & = & \{\,\mathit{Rachael}\,\} \\ V(\text{c}) & = & \mathit{Rachael} & V(\text{loves}) & = & \{< \mathit{Nathan},\, \mathit{Rachael} >\} \end{array}$$

Early MT employed a particular logical language known as first order predicate calculus (FOPC) as a system for the representation of semantic knowledge. FOPC is purely extensional, which means for a given model the correspondence between linguistic symbols and individuals is direct, i.e. FOPC supports a pure MTW perspective. Consequently, lexical symbols such as proper names get their meaning via their capacity to correspond to individuals (e.g. 'Nathan' $\rightarrow$ *Nathan*), and expressions like nouns, verbs, and adjectives (predicates) get their meaning via their capacity to correspond to sets (e.g. 'male' $\equiv$ male(x) $\rightarrow$ $\{Barry; Nathan\}$). These early theories (Lyons 1977, Sabah 1993, Chierchia & McConnell-Ginet 1990) proved extremely limited in their ability to capture natural language semantics, and as such are not considered in the ensuing debate.

Contemporary MT is often described as intensional, and owes much to the ground-breaking work of Montague (1988). In intensional theories lexical symbols obtain their meaning by way of 'intensional concepts' (or simply 'intensions'). These intensions specify the necessary conditions for reference, i.e. they serve to pick out the individual, or set of individuals within the model to which a particular symbol refers. Hence, intensional theories support an integrated MST perspective. However, in practice "people who do model theory of this kind don't really construct set-theoretical models of the universe" (Lakoff 1987, p.178) — when employed within 'understanding' systems (see section 2.3) only the intensional level of knowledge description is utilized. Since intensions are defined purely in terms of symbols, from a practical perspective intensional MT is established upon a MSM foundation.

Meaning postulates (Carnap 1952) are an important component of intensional MT in that they provide a logical device through which intensional definitions (i.e. the contractual relations that hold between predicates, or lexical symbols at all times) are specified, e.g.

$$\forall x[man(x) \rightarrow human(x)] \tag{2.1}$$

This is typical of the sort of meaning postulate a designer might build into an artificial NLP system on the basis of introspection. It states that for all individuals x, if x is a man then x must also be human.

2.2.2 Semantic Primitives

One of the first fundamental departures from MT came in the form of *Componential Analysis* (CA), which is also commonly referred to as decompositional semantics. The inspiration for CA came from work in the field of Anthropology, but was pioneered from an NLP perspective by Katz & Fodor (1963). CA provides a different approach to the description of meaning which rests upon the thesis that words, phrases, etc. can be analysed in terms of a fixed set of more general semantic features, which take the form of symbolic primitives. Hence, CA conforms to the MSM point of view. The use of primitives is motivated by semantic inference (see section 2.3), the view being that one will require fewer, more general, inference rules to act upon the limited set of symbols.

In common with MT, CA typically utilizes a formal language of logic as a means to represent semantic knowledge. A well versed example of lexical CA is that first described by Katz and Fodor which provides the meaning of the predicate *bachelor*(x) :

$$\lambda x\ (\textsc{male}(x) \wedge \neg\textsc{married}(x) \wedge \textsc{adult}(x) \wedge \textsc{human}(x))$$

i.e. *If x is a bachelor, then x is male, x is not married, x is adult, and x is human*

On the surface the example cited above has a striking similarity to a meaning postulate as used in MT. This is in part due to the fact that the primitives that are employed, e.g. MALE(x), all have an obvious relation to lexical entities of the English language. However, these primitives should not be confused with their lexical counterparts, as Lyons notes "MALE, ADULT etc., are held to belong to a set of universal atomic concepts which may or may not be lexicalized in particular languages" (Lyons 1977, p.318).

The notion of primitives also features strongly in Fodor's (Fodor 1975) seminal Language of Thought (LOT) Hypothesis. In this work the author proposes an inner language of the mind, or Mentalese; a rule governed representational system of thought that underlies all other representational systems such as language.

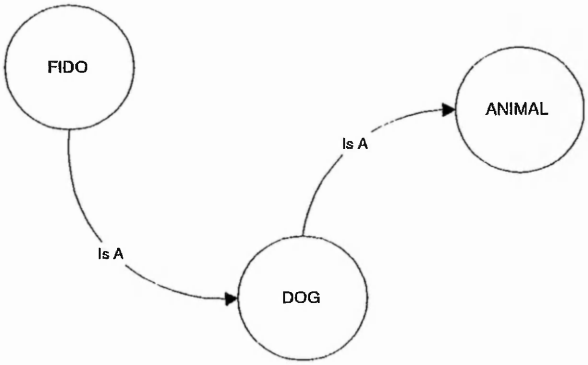

Figure 2.3: A Simple Semantic Network.

2.2.3 Nodes, Arcs, and Labels

It should be clear from the discussions above that MT and CA were developed upon a foundation of classical logic, i.e. systems of symbolic logic provide the substrate for handling semantic knowledge. From the position that such logic based semantic representations were too abstract for use in real natural-language understanding systems, researchers working in the field of AI developed an alternative system of semantic representation known as the *Semantic Network* (SN). Using this formalism the semantic knowledge associated with language elements is mapped onto conceptual structures where each structure is implemented as a network of concepts. A typical network, as shown in figure 2.3, is crafted of three basic components :

- **Nodes** in a network represent concepts of entities, attributes, events, or states.
- **Arcs** in a network, usually called conceptual relations, represent relationships that hold between the concept nodes.
- **Labels** on the arcs specify the relation types.

The study of semantic networks in AI has been quite prolific and the application of such knowledge representation systems has been diverse[2]. In the area of natural language, relational graphs — a specific type of semantic network — have been frequently used. The

[2] A volume edited by Sowa (1991) provides an excellent overview of semantic networks and details many uses of such systems. More recent applications of large scale semantic networks can be found in (Shapiro 2000)

most notable of these are probably Conceptual Dependency (CD) (Schank 1973), Preference Semantics (PS) (Wilks 1975), and Conceptual Graphs (Sowa 1984).

Putting aside their different computational structures, CD and PS are both founded on similar principles to CA: the MSM principle, and the use of semantic primitives[3]. Sowa's approach on the other hand has more in common with intensional MT :

> "Model theory can be adapted to graphs in a concise and elegant way, since the data structure of a model is naturally graph like. Any set of individuals and relations can be represented by a graph with the individuals as the nodes and the relations as the arcs; each arc is labelled with the name of a relation." (Sowa 1987, p.1020)

2.3 Inference, Symbols, and Understanding

A discussion of the intricacies, and of the comparative advantages and disadvantages of MT, CA, and SN is beyond the scope and aims of this chapter[4]. The objective here is simply to indicate the common way in which these theories of meaning support logical *inference* — an ability that has traditionally been equated with that of *understanding*.

Inference is at the heart of symbolic logic. The idea is that one can characterize sound logical arguments by way of general formal rules. In the context of traditional NLP, it is these rules that are thought to embody meaning. For example, consider the following rule, which states 'for all things x, if x is predicated of A then x is also predicated of B'[5]

$$\forall(x) \;\; A(x) \rightarrow B(x)$$

The utility of such general rules does not become obvious until one gives A and B an interpretation, e.g. let A represent 'man,' and B represent 'human.' The rule now reads,

[3] Schank's CD system employs fourteen primitives. This rises to about eighty in Wilks' PS.

[4] Discussions of this kind can be found in (Hirst 1987, Gazdar 1993, Sabah 1993).

[5] Or, in plain English, if one is told that some entity can be validly described as an A, one can conclude that this same entity could also be described as a B.

with the interpretation in place, 'for all things x, if x is a man then x is human' (cf. equation 2.1). Given this rule *a priori*, an NLP system that encounters a proposition such as 'Nathan is a man,' where 'Nathan' takes the place of the entity x, can conclude that Nathan is human. Whilst such a simple conclusion does not instinctively warrant one to suggest that the system can understand natural-language, when many rules are combined the complexity of the system's inferential competence gains stature and there can be, and indeed has been, a strong urge to grant that the system actually understands the language it encounters. For example consider Marconi's description of such a system :

> "From 'There are four elephants in the living-room' our system would infer that there are four large animals in the living-room, that there are four elephants in the house, that there is an even number of elephants in the living-room, ...; it could even infer that the living-room's furniture is likely to be badly spoiled." (Marconi 1996*a*, p.22)

'Natural-language understanding systems' such as those that can summarize portions of text, e.g. FRUMP (Fast Reading Understanding Memory Program) (Schank & Abelson 1977), answer questions in relation to portions of text, or stories, e.g. GENESIS (GENEralising Explanations of Stories Into Schemata) (Mooney 1985), and enter into text based dialog with a person concerning a simple simulated domain of activity, e.g. SHRDLU (Winograd 1972, Winograd 1973) have all operated on the basic principles cited above.

What is worth making quite clear here is the fact that the process of inference (understanding), as described above, is not dependent upon any interpretations (e.g. that A represents 'man'), the rules operate purely on the distinct form of each symbol — this fact is captured in Harnad's description of a symbol system:

> "A symbol system is (1) a set of arbitrary *physical tokens* (scratches on paper, holes on a tape, events in a digital computer, etc.) that are (2) manipulated on the basis of *explicit rules* that are (3) likewise physical tokens and *strings* of tokens. The rule-governed symbol-token manipulation is based (4) purely on the *shape* of the symbol tokens (not their 'meaning'), i.e. it is purely *syntactic*, and consists of (5) *rulefully combining* and recombining symbol tokens. There

> are (6) primitive atomic symbol tokens and (7) *composite* symbol-token strings. The entire system and all its parts - the atomic tokens, the composite tokens, the syntactic manipulations (both actual and possible) and the rules — are all (8) *semantically interpretable*: The syntax can be *systematically* assigned a meaning (e.g. as standing for objects, as describing states of affairs)." (Harnad 1990, p.336)

In the traditional semantic theories described in the preceding section, the arcs, individuals, predicates, postulates, primitives, nodes, names, and networks are all examples of arbitrary physical tokens or strings of such tokens.

2.4 Against the Traditional Notions of Understanding and Reference

Of late, traditional semantic analysis, as put forth above, has become the focus of growing concern. In the discussions that follow a number of related arguments and critiques are presented that capture the essence of these concerns.

2.4.1 Searle, Harnad, and Chinese Symbols

In 1980 Searle (1980) presented a now famous critique — known as the *Chinese Room Argument* (CRA) — in order to discredit claims of natural-language understanding being brandished by the creators of certain artificial language systems, e.g. FRUMP; SHRDLU. Searles thesis contends that any real understanding could ever arise as a consequence of formal manipulation of symbols, i.e. logical inference. To clarify his position Searle devised an intuitive thought experiment, which can be roughly reconstructed as follows:

> Imagine that a native English-speaking person who knows nothing of the Chinese language is locked in a room which contains a book of rules that is written in English and a big pile of Chinese symbols, which to the person are all meaningless squiggles. Now suppose that they are passed (e.g. under the door) a sequenced batch of Chinese symbols. Using the English rules they are able

> to correlate and manipulate the symbols. The rules specify the manipulations of the symbols purely formally, i.e. the symbols are identified solely by their different shapes, for example a rule could say: 'if you receive a symbol like this ( ) you may replace it with a symbol ( )'. Finally, the person is passed (under the door) a second batch of symbols. Other rules specify, again formally, how — on the basis of all the symbols to which they now have access — they should return (under the door) certain symbols from the pile.

As Searle points out, the batches of symbols could, unknown to the person in the room, in fact be (1) a story, and (2) questions relating to the story (all in Chinese written language). In addition, the symbols that are returned could in fact be answers to the various questions. Hence, Chinese-speaking observers outside the room may very well conclude that the person in the room understands Chinese, even though it is obvious that he/she does not. Searle concludes that traditional inferentially competent NLP systems, that operate in a purely formal manner (i.e. like the person in the room), cannot be said to 'understand' what they are doing or processing, noting that this is because the symbols lack intentionality.

The intentionality dilemma initially pointed out by Searle was later picked up by Harnad (1990) and dubbed as the *Symbol Grounding Problem* (SGP). Harnad argues emphatically that traditional 'understanding' systems do not understand because they do not know the meaning of *any* of the symbols that they employ:

> "The symbols and the symbol manipulation, being all based on shape rather than meaning, are systematically *interpretable* as having meaning ... But the interpretation will not be *intrinsic* to the symbol system itself." (Harnad 1990, p.338)

The essence of the SGP is that a meaningless (i.e. uninterpreted) linguistic symbol, cannot be made meaningful by being connected to more uninterpreted symbols (cf. the MSM perspective of meaning). To corroborate this claim, Harnad invites the reader (who he assumes to be a non-Chinese speaker), to envisage the task of learning Chinese with only a Chinese/Chinese dictionary at ones disposal. The author notes that a 'trip through the dictionary would amount to a merry-go-round, passing endlessly from one meaningless symbol

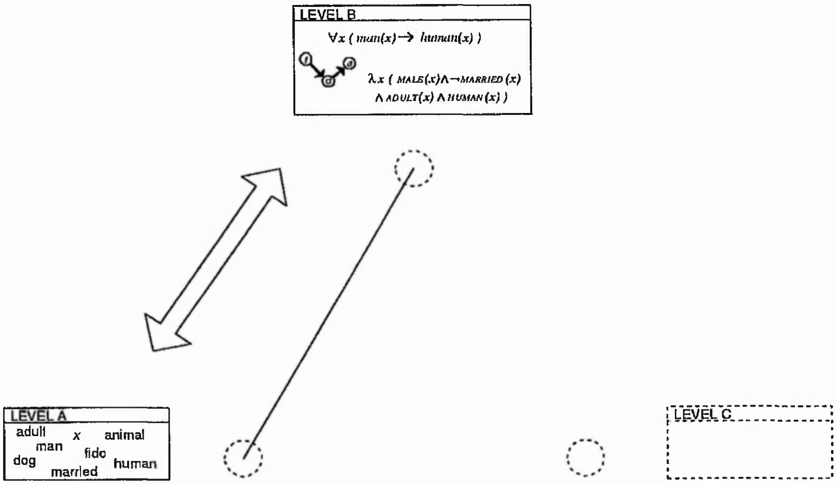

Figure 2.4: In symbol systems (e.g. MT, CA, & SN) the meaning of any particular symbol that resides within the system (at level A) is determined through its relations with other symbols in the system (defined at level B). The symbol system has no connection to the world (level C).

or symbol-string (the definiens) to another (the definiendum), never coming to a halt on what anything meant' (Harnad 1990, p.339).

The SGP (and as a consequence, the CRA) prevails, no matter how many connections and nodes one builds into a semantic network, or how many meaning postulates one writes, or how large ones rule book is, or how many definitions ones Chinese dictionary contains (see also (Partee 1981)). To escape the SGP and make the symbols meaningful from the systems perspective, one must address the issue of how symbols can be grounded in something other than just more meaningless symbols. MT, CA, and SN must provide an explanation of how the symbols (possibly just the primitives) within their internal computing languages are connected[6] to those entities which they are supposed to symbolize.

[6] In the context of these discussions the term 'connected' is synonymous with, and could be replaced by, the following terms: 'grounded' (Harnad 1990), 'causally linked' (Fodor 1980), 'causally related,' or 'hooked' (Jackson & Sharkey 1996).

Figure 2.4 has been developed by this author to capture the essence of Searles and Harnads[7] arguments from the graphical perspective of the TOM. The rules and representations that have been created on the basis of introspection and set within the languages of traditional semantic theories for the purpose of inference etc. all exist at the conceptual level (B). The NLP systems that employ these theories have no access to the entity, or world level (C) and thus, to put it bluntly, they don't know what they are talking about.

[7] The issues highlighted in figure 2.4 which are at the core of Searles and Harnads critiques have also been addressed by several other authors. Variations include Lloyds debate concerning meaning without ostension, Chalmer's problem of *representation grounding* (Chalmers 1992), Sharkey and Jackson's *internalist trap* (Sharkey & Jackson 1994), Bickard and Terveen's program of *semantic circularity* (Bickhard & Terveen 1995), and Marconi's discussions concerning *referential incompetence* (Marconi 1996*b*, Marconi 1996*a*).

2.4.2 Reference, Lakoff, and the Nature of Categories

In light of the critiques discussed above, traditionalists concede that an autonomous system of symbols cannot in itself generate meaning, i.e. a pure MSM perspective is flawed. They generally agree that to create an artificial system to which understanding could be truly attributed, one requires an account of reference[8] — how symbols (some at the very least) are connected to the world; the integration of level C of the TOM.

The account of reference that is favoured by the traditionalists is closely allied to an MTW account of meaning. They believe that their rules, relations, symbols etc. can easily be given an intrinsic interpretation (to the system) because they are internal representations of an objectively existing external reality, i.e. they actually embody the way the world is (cf. a 'Mirror of Nature' (Rorty 1980, Lakoff 1987)). Since what is inside the system is isomorphic with what is outside the system, the defenders of tradition suggest that only simple peripheral devices should be required in order that a system "see" the world of entities to which its symbols refer :

> "[Fodor] posits a passive mechanism of perceptual analysis ... Demons each sensitive to a single physical parameter shriek *yes* or *no* depending on whether a hypothesis is present or absent in the environment. These demons activate innate elementary concepts which, once properly combined, are used to reason formally about the world." (MacDorman 1995)

And on the presumption that these symbol-world connections are so trivial (i.e. 1:1) assume that they are of little academic, or practical interest :

> "While connections to the world and symbol manipulations are both presumably necessary for intentional processes, there is no reason (so far) to believe that the former provide a theoretical domain for science." (Fodor 1980, p.431)

However, numerous findings which have recently been drawn together by Lakoff (1987)

[8] For example, in response to Searle's CRA, Fodor points to a need for 'causal linkages between the symbols that the device manipulates and things in the world' (Fodor 1980, p.431). See also (Pylyshyn 1980, Winograd 1980).

suggest that there *is* now reason to believe that (1) connections to the world are of great interest, (2) that passive perception will just not do, and consequently (3) that traditionalists have grossly underestimated the difficulty of picking out the entities that symbols refer to. In other words, their theoretical solution "trivializes the symbol grounding problem" (Harnad 1990, p.340). In a comprehensive volume Lakoff collates a wealth of diverse empirical data and logical argument all of which refutes an MTW (which Lakoff calls 'objectivist') reading of reference. The main evidence is as follows :

- **The Nature of Categories**: To successfully map between symbols and the world, an objectivist theory requires that the categories of things (in the outside world) be Aristotelian, i.e. that category membership is an all-or-nothing phenomenon characterized by a set of necessary and sufficient conditions. However, empirical research on human categorization across a broad spectrum of disciplines has provided contrary evidence that suggests category membership is of much greater complexity comprising phenomena such as fuzziness and prototypicality (Wittgenstein 1953, Rosch 1973, Lakoff 1987). For example, a cat that has been in an accident and has lost a LEG, its TAIL, an EYE, and its WHISKERS is still a 'cat'.

- **An Objective World**: Lakoff also argues at length, with support from empirical findings, against the notion that certain categories exist in the *outside* world at all. This is qualified by pointing to, amongst other things, colour categories:

  > "Wavelengths of light exist in a world external to human beings; color categories do not. The fact that we categorize different wavelengths as being in the same category partly depends on human physiology — on the cones in the retina and the neural pathways between the eye and the brain. Colors arise from our interaction with the world ... Colors are categories of the mind that do not exist objectively in the world exclusive of seeing beings." (Lakoff 1987, p.198)

- **Putnam's Proof**: Lakoff also devotes a full chapter of his volume to a complex logical proof (which is beyond the scope of this thesis) devised by Putnam (1981) that clearly demonstrates fundamental flaws in the MTW account of meaning.

2.4.3 The Problems in Perspective

In summary model-theory, componential analysis and semantic networks, are all inadequate as *complete* theories of meaning because they all lack an appropriate account of how their symbols refer (i.e. how they are grounded). Without such an account all of the symbols are meaningless to the system and thus the system cannot be said to understand.

However, these problems do not preclude the use of traditional MT, CA, and SN. One must simply be aware of certain constraints and realize that NLP systems that employ these formalisms are simulating understanding rather than actually understanding. The knowledge that is encoded within these traditional frameworks *is* meaningful *but* only to those outside of the system, it is *not* intrinsic to the system[9]. Hence, NLP systems that utilize such knowledge are simply 'dumb' intermediaries between a designer (who understands the referents of his symbols) and an end user (who is assumed, by the designer, to understand the referents of the designers symbols with a correspondence to the designers understanding). Accordingly, if designers can amass enough consistent knowledge (formal rules, relations etc.) they may well be able — and it would certainly be no mean feat (see the 'Frame Problem', section 1.2.3) — to empower the person in the Chinese room to actually fool those that are outside[10], i.e. they may be able to create artificial systems that can pass the Turing Test (Turing 1950). And without a doubt such systems would prove to be extremely useful in many areas of application.

The area of application that is under investigation in this thesis — natural-language communication between humans and embodied artificial systems which operate within a physical environment (e.g. robotic systems) — is, however, one domain where simulated understanding does not suffice as a substitute for the real thing. Embodied systems with sensorimotor

[9] Alternatively, one might say that MT, CA, and SN provide a partial theory of meaning, i.e. there is certainly meaning within a Chinese/Chinese dictionary, however, the 'relational' type of meaning that there is does seem dependent upon a more fundamental 'referential' type of meaning.

[10] However, it must be noted that designers may be foolhardy to disregard the issues raised above. As Harnad (Harnad 1990, p.340) suggests "the fact that our own symbols do have intrinsic meaning whereas the computer's do not, and the fact that we can do things that the computer so far cannot, may be indications that even in AI there are performance gains to be made (especially in robotics and machine vision) from endeavoring to ground symbols" (see also (Glenburg 1997)).

peripheries must be able to apply their words *vis-à-vis* the real world, i.e. their symbols *must* refer. If such systems are commanded to 'pick up the blue cube', they must know what the symbols 'blue', 'cube' etc. refer to within their immediate environment. Hence, when working with such systems the issues raised above are *fundamentally important.*

2.5 Rethinking Semantics and Meaning

In this, the closing section of the chapter, discussions are focussed towards a number of associated ideas, claims, and proposals which have been volunteered by various authors in relation to a *complete* theory of meaning, i.e. one that incorporates an adequate account of how symbols refer (are grounded). The objective of this section is to draw upon these various sources in order to provide a single coherent perspective of a *non*-traditional contemporary semantic methodology. To achieve this goal Ogden & Richards (1923) graphical tool of description, the triangle of meaning, has been adapted and extended.

2.5.1 A New Triangle of Meaning

This thesis proposes a new general acccount of lexical meaning which is embodied in figure 2.5. This adapted TOM provides a novel, and perhaps more intuitive means of interpretation for a number of essentially analogous semantic theories which have been proposed by various authors (see, for example, (Harnad 1990, Dorffner 1992, Gärdenfors 1995, Jackson & Sharkey 1996)). For the purpose of the discussions that follow throughout this thesis the terminology 'embodied semantics' will be used to differentiate a theory of meaning that fits with the adapted TOM from the traditional theories detailed above, i.e. MTW, MSM, and MST. Furthermore, the acronym ENLP (Embodied NLP) will serve as a description for natural-language processing systems that are built upon a foundation of embodied semantics, and as a consequence possess the correct structure to harbour a meaningful lexicon.

The adapted TOM captures *four* key levels of description that are integral to a theory of embodied semantics: the world level (D), the sensory level (C), the conceptual level (B), and the symbolic level (A). The ensuing sections consider each of these levels in detail.

The World Level

One of the fundamental views that is central to embodied semantics, and one which radically differentiates this type of theory from those that have been described above, is that the world — as it is 'out there,' external to language using systems — is not something that is neatly carved into nice objectively existing chunks (objects, events, etc.). Strictly speaking, from this alternative perspective, the objects that humans "see" do not exist as single bounded entities independent of humans (cf. (Jackson & Sharkey 1996)). Lakoff lends support to this claim:

> "Take, for example, the chair I am sitting on. It exists. If it didn't, I would have fallen on the floor. But that chair can be viewed *correctly* in many ways. From the molecular point of view, it is an enormous collection of molecules and not a single undifferentiated bounded entity. From the point of view of wave equations in physics, there is no chair, but only wave forms. From a human point of view, it is a single object." (Lakoff 1987, p.262)

Jackson & Sharkey (1996) have described what lies outside of the language using systems as a "rich and heterogeneous soup of electro-magnetic, chemical, and mechanical energies." However, one should be wary of falling into the trap of using such 'fundamental' words to describe the world outside of an embodied system[11]. Following from the chair example one could rightly argue that a mechanical energy is simply another human construct – one interpretation of that which surrounds one. Here, as in Chandler et al. (1999), the world level will simply be considered as an unknown.

Hence, it should be clear from this new recognition of reality that symbols cannot refer directly to things in the world (e.g. as in MTW and MST) but rather, as a logical consequence, reference must be directed toward something that is internal to the language using system.

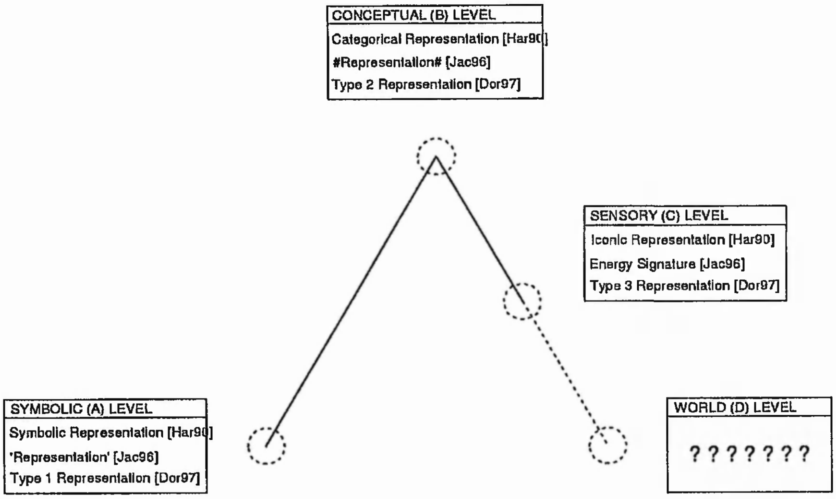
CONCEPTUAL (B) LEVEL
Categorical Representation [Har90]
#Representation# [Jac96]
Type 2 Representation [Dor97]
SENSORY (C) LEVEL
Iconic Representation [Har90]
Energy Signature [Jac96]
Type 3 Representation [Dor97]
SYMBOLIC (A) LEVEL
Symbolic Representation [Har90]
'Representation' [Jac96]
Type 1 Representation [Dor97]
WORLD (D) LEVEL
???????

Figure 2.5: A new triangle of meaning

The Sensory Level

Although meaning is taken to be internal to the system, an embodied semantic methodology does not ignore the world level (cf. the MSM philosophy). From a human perspective, the fact that one "sees" a chair (and thus knows to what the word 'chair' refers) is without doubt, a function of something that derives from ones external environment. Accordingly, peripheral sensory devices — whose function it is to measure and then convert these unknowns to some usable internal representational format — are considered to be absolutely essential within this new semantic methodology.

Level C within the new TOM recognizes that a certain state within the world will give rise to a specific response amongst a system's sensory devices. Given the causal nature of the connection, such a response is obviously non-arbitrary. Hence, the sensory level of description is inherently *non-symbolic.* These non-symbolic sensory responses have been given different labels by different authors: *Iconic Representations* (Harnad 1990); *Energy Signatures* (Jackson & Sharkey 1996); *Type-3 Representations* (Dorffner 1997).

[11] Many thanks to Alex Reigler (personal communication) for pointing this out.

The Conceptual Level

Whilst sensory devices provide the all important connections to the world, they alone do not possess sufficient computational power to satisfy an appropriate account of reference (i.e. generate meaning). Since entities, and categories of entities do not exist in the external environment *per se*, sensory devices cannot simply pick them out. In a theory of embodied semantics the process of *categorization* is proposed as the final piece in the puzzle of meaning.

> "Imagine that on a particular morning, while you are brushing your teeth, your ability to perceive similarities between objects slowly dissipates. You first notice that the red object in your hand behaves strangely. Each time your hand moves, the object looks completely different. Then you look in the mirror, at the object you were sure was your face a few minutes ago. You tilt your head to the left, then to the right, but each time your own reflection in the mirror looks like a different face. *Toothbrush* and *face* as permanent entities have just vanished from your mental life. Your mind does not "see" through your eyes anymore. As you contemplate this continuous flow of unrelated experiences, you suddenly realize you are less and less able to forget a difference, to generalize, to abstract ... to think." (Schyns & Rodet 1995, p.234)

The passage above is cited because it succinctly captures just why sensory responses, which result from transduction at level C, are not immediately useful in and of themselves. The additional act of categorization is required to segment, or carve the mass of undifferentiated environmental stimuli into mental representations, or concepts, on the basis of similarities (e.g. distinctive, or invariant features). It is these representations at level B within the new TOM — christened by Harnad (1990) as *Categorical Representations*, by Jackson & Sharkey (1996) as *#Representations#*, and by Dorffner (1997) as *Type-2 Representations* — which essentially *are* the objects, events, etc. that lexical symbols symbolize. Since level B representations are also directly coupled to sensory input, they are non- arbitrary, and consequently non-symbolic in nature.

The Symbolic Level

The final level of description within this new semantic formalism is a place where the arbitrary symbols and manipulations thereof can exist happily — supported by the other levels.

Once some set of lexical symbols has been appropriately grounded, the symbolic level can come into its own. A frequently cited hypothetical example (originating from Harnad (1990)) demonstrates how this might work. Suppose that the lexical symbols 'horse' and 'stripes' are grounded through levels B and C as specified above. Now the meaning of an unknown symbol 'zebra' can be specified by way of a purely symbolic proposition, e.g. a 'zebra' is a 'horse' with 'stripes.' The symbol 'zebra' inherits a grounding. As Harnad notes: "In principle, someone who had never seen a zebra (but had seen and learned to identify horses and stripes) could identify a zebra on first acquaintance armed with this symbolic representation alone (plus the nonsymbolic – iconic and categorical – representations of horses and stripes that ground it)." (Harnad 1990, p.343)

2.5.2 A Brief Synopsis

In summary, a semantic formalism (embodied semantics) has been described that provides a theoretical starting point from which to devise natural-language understanding systems that are *referentially* as well as inferentially competent. The core aspects of this alternative formalism are as follows:

- Lexical symbols (i.e. words) refer to, and hence are meaningful with respect to, internal concepts.
- Internal concepts are mental representations created by, and for, a particular individual, hence meaning, and ones knowledge and understanding of the world is subjective rather than objective.
- The process of concept creation centers around the transduction and categorization of environmental stimuli.

The chapter that follows considers the practical implementation of such a semantic theory. First it addresses how connectionist models — systems that employ 'spatial' rather than symbolic (or logical) reasoning (cf. Jackson & Sharkey (1996) — provide a suitable computational framework for such a theory). This is followed by a discussion of how various connectionist models have actually been employed in ENLP systems.

Chapter 3

Connectionist Models of Lexical Acquisition and Grounding

The significance of the arguments put forth in chapter 2 has resulted in the development of a 'new wave' of applied language systems in the last ten or so years, all of which adhere in some respect to the new theoretical framework previously described. In the second part of this chapter the most important of these models are introduced, described, and then critically assessed with respect to a number of important criteria. However, as an important precursor to these discussions, connectionism and its application *vis-à-vis* the task of acquiring a grounded lexicon is first given some consideration.

3.1 Connectionism

All of the language systems reviewed in section 3.2, with the exception of one, employ at their core, connectionist computing devices. In order to provide a suitable platform from which to discuss these language systems it seems appropriate to first consider the nature of such devices. However, the purpose of this section is not to provide an exhaustive historical account of connectionism, nor is it to detail the individual mechanics of the myriad of contemporary connectionist computing devices. The purpose here is to focus on the general computational characteristics shared by these devices, and as a consequence to :

1. Illustrate how connectionism naturally satisfies the representational requirements of an embodied semantic methodology, and accordingly, provide an instructive connectionist interpretation of the theoretical ideas set out in the final section of chapter 2.

2. Show that connectionism naturally provides a developmental framework for language processing. Semantic knowledge can be *acquired* rather than explicitly programmed (cf. the DC approach; section 1.2.3).

3.1.1 Computing with Connectionist Models

Connectionist models[1] (CMs) are characterized by two general features: (1) They consist of separate layers of interconnected simple homogeneous computing elements, or units (see, for example, figure 3.1a), and (2) they employ recursive rules for updating the strengths of the connections (often termed as 'weights') during learning.

Connectionist models are in essence pattern classifying devices. They are able to group together different sets of input patterns into various classes or categories (the learning function) — at a more abstract level one could say that CMs induce a set of input-output rules. Subsequently, given a new and perhaps novel input pattern, a CM is able to predict to which output class or category the input belongs (the prediction function). What follows is a brief account of the input to, and the internal structure of a CM. This provides a basis from which to describe in greater detail the general mechanics of learning and prediction.

The Input to a CM

The input to a CM is an n-dimensional vector, often referred to as a feature vector. Each slot of the feature vector encodes the degree of presence (or equivalently, absence) of a particular property by way of a bounded range of values. For example, using the numerical

[1] Connectionist models are also often referred to as 'Neural Networks,' or 'Artificial Neural Networks.' Whilst the latter of these alternative descriptions is better than the first, both are still somewhat misleading. Connectionist models are merely inspired by brain-like computation, and most researchers now realize the vast chasm that exists between the make up of these models and the neural networks that nature's creatures employ. As Hanson and Burr (Hanson & Burr 1990, p.472) so aptly comment, connectionism is "something that looks like a cartoon brain with cartoon neurons, connected by cartoon synapses."

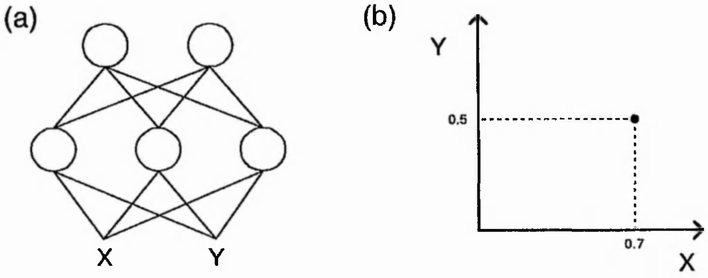

Figure 3.1: (a) Shows a simple feed-forward multi-layer CM with inputs X and Y, a layer consisting of three hidden units, and a layer of two output units. (b) Depicts the 2-dimensional feature space for this CM. An input of {X = 0.7, Y = 0.5} produces the point in feature space that is shown.

range [0,1], if the n^{th} property in question is fully present then the n^{th} dimension of the vector will have a value of 1. Alternatively, if the property is only partially present, it may have a value of 0.5, and so on. The dimensionality of the input, and the numerical range of each dimension effect an input, or feature space, i.e. a continuous n-dimensional vector space that subsumes all possible input vectors. Figure 3.1b shows the 2-dimensional feature space affiliated with the CM shown in figure 3.1a.

The Internal Structure of a CM

Feature vectors are propagated to a layer of output units by way of the relevant connections. Each connection is weighted — hence, the description 'weights.' The weighting effectively regulates the proportion of the input that each unit receives. Each unit integrates all of the input that it receives by way of a fan-in, or activation function. From a geometric perspective these fan-in functions realize the formation of partitions within the feature space called decision boundaries. These decision boundaries are at the heart of a CMs classification, or categorization abilities, in that they carve the feature space into various regions each of which can be employed to enclose a set of points.

The complexity of a CMs categorization abilities can be governed using two distinct methods. One method centers around the complexity of the fan-in function. As shown in figures 3.2a-c different functions give rise to differing types of boundary and hence to different forms of categorical region within a feature space. Alternatively, another way to increase categorization complexity is to use extra layers of 'hidden' units (Hanson & Burr 1990,

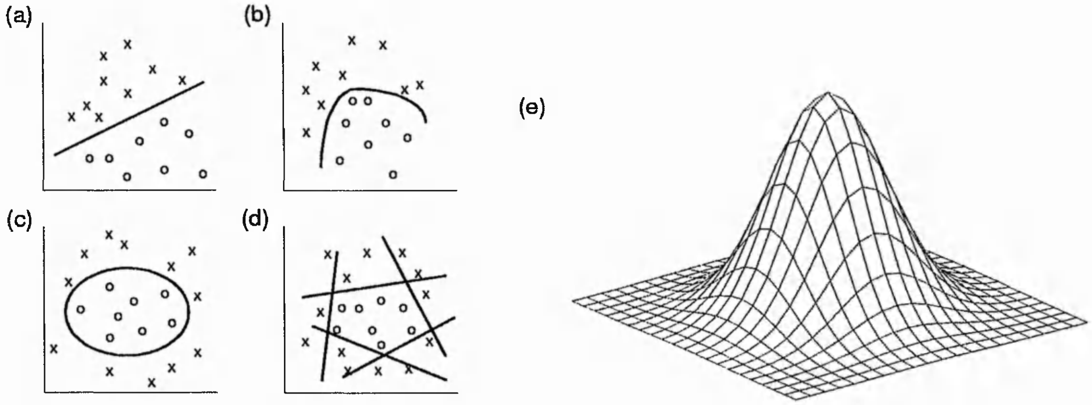

Figure 3.2: (a-c) Show how three different fan-in functions (linear, quadratic, and spherical) create contrasting boundary forms in a 2-d feature space. (d) Shows five linear boundaries being used in conjunction to create a categorical region. The job of the learning algorithm is to adapt the position and dimensions of the boundaries so as to create 'categorical' regions that correctly separate the o's from the x's. The concurrent manipulation of multiple boundaries to create a single region generally requires the use of a supervised learning process (see text). (e) Shows a boundary in a 2D feature space created by a Gaussian fan-in function (which is a special case of a spherical function). In this case the third dimension (in the z plane) provides a measure of goodness.

p.476) between the input and output layers, as is the case in popular multi-layer perceptron (MLP) models (Rumelhart et al. 1986). Each hidden unit effects a simple boundary. The hidden to output layer dynamics make it possible for the CM to combine these boundaries in order to create potentially complex regions within the feature space as is graphically illustrated in figure 3.2d.

Learning in a CM (the acquisition of knowledge)

The fan-in function is non-adaptable, and hence the general form of a particular boundary in feature space is fixed. However, the weights that are associated with a fan-in function are adaptable. It is these 'free' parameters that determine the specific placement, and the dimensions of a given boundary form within the feature space. Learning within a CM concerns the manipulation of the model's many weights, and hence, the position and dimensions of numerous boundaries. The 'training' algorithms that enable this function generally come in two flavors :

- **Unsupervised** : These algorithms discover similarities between input vectors without feedback, i.e. the CM has no external 'teaching' signal informing it to which class (within which bounded region) a given input vector belongs. If an input vector i falls within the limits of a classification region that is affiliated with an output unit j, then i may be considered as a member of the class C_j. Consequently, i may influence the size and position of this region — the weights between the input and unit j are updated to take into account the new member.

- **Supervised** : These algorithms employ feedback. Input-target pairs are presented to the CM, e.g. $\{i, C_k\}$ if input i is a member of class C_k. An error measure defines how far outside the correct region of feature space the input i is. Weight adaptation seeks to minimize this error.

Whichever learning formalism is used, the adaptation of decision boundaries enables the CM to geometrically separate, given the context of the feature space, certain sets of input vectors from others, e.g. the o's and x's in figure 3.2. What distinguishes the two formalisms is their respective abilities to craft the shape of categorical regions. On this front the supervised process is generally considered to be more powerful; indeed it has been theoretically shown that a multi-layer CM employing supervised learning can learn any non-contradictory input to output mapping (Cybenko 1989, Hornik et al. 1989).

Prediction in a CM (the application of acquired knowledge)

One of the most enticing properties of CMs is that they don't just learn the information that they are provided in a rote manner. They possess the correct computational apparatus to actually learn *from* the information that they are provided. CMs naturally undertake generalization. The categorical regions in feature space that result from learning not only separate the various input vectors that were used in the learning process. Given the continuous nature of the feature space, each region actually separates a potentially infinite number of points in space, each of which is associated with a possible input vector. Hence, on the basis of which particular region a novel input vector falls within a CM, it can hypothesize *vis-à-vis* which output class the input belongs to. Given the nature of the n-dimensional spaces involved this process of prediction has been described as spatial reasoning (van Gelder

1991, Sharkey & Jackson 1994).

3.1.2 Connectionism and Embodied Semantics

To recap from the discussions presented in section 2.5, an embodied language system must support three levels of representation: sensory, conceptual, and symbolic[2]. CMs with their vector based input, spatial representational substratum, and categorical output, suitably accommodate these requirements and thus provide a suitable mechanism for the implementation of ENLP systems. The ensuing sections address this suitability in greater depth.

Sensory Vector-based Input

Sensor based readings of environmental stimuli are generally multi-dimensional. Consider the human visual system. Whilst at a macro level one might consider an eye as a single sensory device, at a micro level one can observe that each human retina is in fact composed from a multitude of smaller visual sensors (e.g. cones, rods, etc.), each of which may respond differently to an incoming stimuli (Marr 1982). Likewise, consider a common artificial visual sensor, viz. the CCD camera — as frequently employed in robotic systems. The camera takes a reading of the outside environment by way of an i by j matrix of charge coupled sensory devices. As before, each device may respond differently with respect to a constant visual stimuli.

Given a multi-dimensional interpretation, one can see that sensory based inputs have an obvious correlation to the n-dimensional vector type inputs processed by CMs. Hence, CMs can process the required sensory data directly, e.g. each elementary sensing device within a CCD camera has the task of measuring one 'feature' of a visual input vector that can be propagated to a CMs input units and from there onward to the output units.

[2] The world level is of course not a level of representation. It is simply a level of description that accounts for the phenomena that exist 'out there' in the world independent of any language system.

Concepts in a Conceptual Space

At the conceptual level one requires representations that capture the common characteristics of the entities or categories of entities, about which one wishes to communicate. As discussed previously, sensory input vectors will not suffice to directly fix reference for a particular lexical symbol because the external world is dynamic and in constant metamorphosis (cf. Dorffner's claim that "probably no two patterns [sensory] at different times are ever the same, even if they stem from the same object" (Dorffner 1992, p.279).

Given this requirement one can immediately see the applicability of CMs. The learning process serves to separate a number of o's (e.g., visual input vectors arising from observations of cats), and x's (... observations of dogs). After successful separation, the resulting categorical regions capture the features, or relations between features that are invariant within the set of cat vectors (i.e. a cats common visual characteristics), and likewise for dog vectors. Hence, a CM's categorical regions posses the appropriate representational qualities to satisfy basic conceptual level obligations and one may interpret these regions within feature space as concepts situated within a conceptual space.

Applying Output Labels

Finally, satisfying symbolic requirements, a CM's output nodes – each relating to a different categorical region within the feature space – can be assigned an arbitrary label (lexical symbol), e.g. following from above one node would be assigned 'cat,' the other 'dog.'

Hence, looking at CMs from a semantic perspective, one has lexical symbols (arbitrarily named output nodes) that refer to, and hence are meaningful with respect to, internal concepts (that are directly connected to n-dimensional regions of feature space). The internal concepts are triggered by external stimuli, i.e. by readings of environmental energies at ones sensory peripheries (different feature vectors will fall within different categorical regions).

3.1.3 CMs and Lexical Development

It has been shown above exactly how CMs embrace the embodied semantic methodology that is vital if the lexical symbols at the heart of any language system are to be intrinsically meaningful. Thus, as an implementational platform, it would seem that CMs are adequately suited to the task of grounding a lexicon. However, reiterating from chapter 1, the overall objective of the work presented in this thesis is to create a system that can *acquire* grounded lexical symbols as and when needed, a system that can develop autonomously just as a child does during its early years. The CMs as described above go part way toward this goal in that they support the acquisition of sensory based semantic knowledge, i.e. the shape and position of the categorical regions in feature space are not explicitly programmed, they are adapted with respect to certain input data using a learning process. However, the lexical symbols are not acquired, they are simply associated with output nodes in an arbitrary fashion by the model's designer. The model as described above does not address the issue of how the lexical entities themselves emerge.

The creation of an autonomous system requires a recognition of the similarities between lexical and non-lexical entities — given an embodied framework. In the same way that the recognition of an object's existence arises from sensory stimulation and categorization, so does a word's existence[3]. The rudimentary sensory input from which a working lexicon is acquired may be acoustic (sound patterns resulting from speech), visual (the squiggles that constitute written words), or even haptic (patterns of bumps in Braille).

Given this sensory based interpretation of lexical symbols one can make an extended claim about basic language development, viz. a lexicon and its semantics grows from a historical coupling of sensory signals. Consequently, one can make an associated claim that ostension[4] gives rise to such a developmental process (cf. (Harnad et al. 1994)). Guided by these principles figure 3.3 portrays a generic lexical acquisition and grounding architecture which comprises two basic CMs fused to form a larger structured model. The architecture

[3] At least this is the case if one observes a learning-theoretic *constructivist* (see, e.g. (Maturana & Varela 1980, Peschl 1992, Quartz & Sejnowski 1997)) perspective of cognitive development rather than a *nativist* viewpoint (e.g., (Chomsky 1980, Pinker 1994)) in which mental entities such as lexical symbols are assumed to reveal themselves according to some predefined congenital time-scale, and learning based upon environmental interaction plays no part. Karmiloff-Smith provides an interesting survey of this philosophical terrain in a recent volume (Karmiloff-Smith 1992).

[4] Ostension is the process by which entities are pointed out and named. For example, an adult may in the presence of a young child gesture toward an animal in a field and say 'look at the horse,' or 'can you see the horse,' or even 'say horse.' These are all ostensive descriptions.

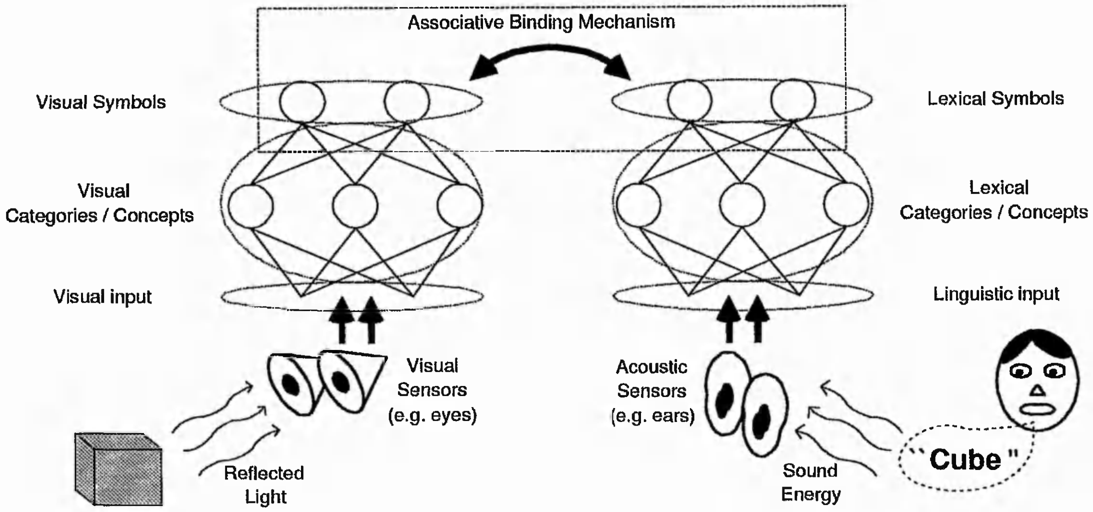
Associative Binding Mechanism
Visual Symbols
Visual Categories / Concepts
Visual input
Lexical Symbols
Lexical Categories / Concepts
Linguistic input
Visual Sensors (e.g. eyes)
Acoustic Sensors (e.g. ears)
Reflected Light
Sound Energy
"Cube"

Figure 3.3: Generic lexical acquisition and grounding architecture. A modular structured CM comprising two sensory input channels, in this case acoustic and visual.

facilitates:

1. The acquisition of sensory derived non-linguistic knowledge, e.g. object concepts derived from visual input.

2. The acquisition of sensory derived linguistic knowledge, e.g. lexical symbols derived from acoustic input.

3. The realization of causal links between concepts/symbols that co-occur on a regular basis, e.g. the creation of a link between an object concept and the word "cube" (see figure).

This generic LAG architecture is at the heart of the novel lexical acquisition and grounding research described later in this thesis, and its general form is evident in a number of the applied language models that are reviewed below (cf. (Prem 1994)).

3.2 Applied ENLP Models

Since 1990 the embodied semantic philosophy has started to take hold, and indeed thrive in the arena of language research. Over recent years an increasing number of language systems have been developed that are inspired and build upon (at least partially) embodied semantic principles. The closing sections of this chapter detail and critically appraise a number of these ENLP models in a chronological order[5]. Before commencing this review process it is instructive to first identify what one might expect from a 'model' model, this can then be used as a gauge by which to assess the capabilities of each model in turn.

3.2.1 A 'Model' Model

Whilst all of the work reviewed in this chapter is related to the novel work presented later (all being founded on similar theoretical foundations), the rationale each author has had for undertaking their work has varied, e.g. in one case an ENLP inspired model is employed in an attempt to explain certain characteristics of language development in children. In the case of the work in this thesis the rationale is perhaps best described as application-oriented, i.e. the aim is not simply to undertake 'pure' language research, it is toward the creation of *practical* systems that can be *applied* in realistic environments to provide an interface enabling natural-language interaction between humans and robotic systems. Given this rationale there are a number of important criteria that one might deem to be fundamentally important, the sort of properties that an ideal model might exhibit:

Learning Properties:

- *Autonomy* – Learning should be autonomous in the sense that usable lexical semantic knowledge should develop naturally as a direct result of the catagorisation, combination, and assossiation of incoming data captured through sensory peripheries. Any

[5] Unfortunately space constraints prevent an all embracing review of related language systems. Those that are reviewed have been chosen because they highlight certain important issues. For reference, further interesting and related systems that are not discussed below include those created by Henis and Levinson (Henis & Levinson 1995) (see also (Gorin 1995)), by Reiger (Feldman et al. 1996), and by Gasser and Smith (Gasser & Smith 1993).

model that requires a designer to continually intervene in order to artificially 'force' data into the model is at a compromise.

- *Flexibility* – Furthermore, learning should be flexible in the sense that the sort of patterns / data that can be learnt are not overly constrained. The need for such flexibility is supported empirically by work presented in chapter 5.
- *Adaptable & Scaleable* – The knowledge acquired through learning should be both adaptable and scaleable. As in the human case, language should develop as and when required in order to communicate an ever growing understanding of an operating environment. One should also recognize the need for a model to be able to integrate many different sensory channels (multi-modality) in order to build realistic lexicons, e.g. one can apply the word 'orange' with respect to visual input, haptic (tactile) input, and taste input. This is a key topic in chapter 7. Furthermore, there is the issue of temporal processing. Verbs such as 'run' can never be meaningful with respect to discrete slices of sensory information.
- *Performance* – A model should be able to acquire knowledge at a reasonable rate, e.g. it should be able to learn new words within a few presentations (a short space of time), just as an adult human does. In addition, to be of any practical application, an ideal model should rely upon affordable computing resourses.

Prediction/Recall Properties:

For a model to be of any use it must be able to effectively recall the knowledge that it has acquired. This can be tested by way of two abilities: description, and identification:

- *Description* – When non-linguistic stimuli are presented alone, a system should activate internally the appropriate lexical symbol(s), i.e. the word(s) that correctly describe that stimuli to an external observer (figure 3.4a).
- *Identification* – When linguistic stimuli are presented alone, a system should activate internally the appropriate non-lexical symbol(s), i.e. it should be able to *imagine* the concept (e.g. an object) to which the input stimuli refers (figure 3.4b).

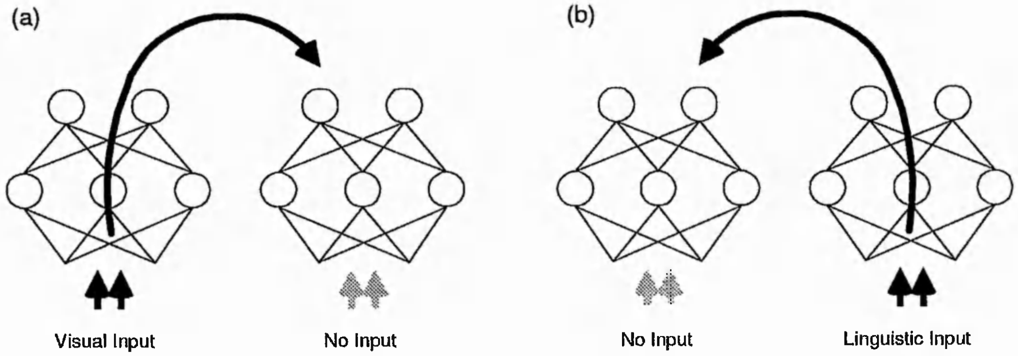

Figure 3.4: Description and identification in a generic LAG architecture: (a) Description – Non-linguistic sensory input that enters the left hand side of the model activates a lexical symbol in the right hand side of the model. (b) Identification – Linguistic sensory input triggers a non-lexical symbol in left hand side of model (the model imagines).

These functions are central to the rest of the investigations detailed in this thesis, since they represent the ability to bridge the gap between words and the world, i.e. they provide some evidence of 'understanding.'

Performance is important in recall just as it is in learning. A model should be able to recall its knowledge (i.e. describe, or identify) in a reasonable period of time. For example, if directed to 'pick up an object' one would hope that a robot would understand and act upon the instruction quickly.

With the make-up of a perfect system now specified, the remainder of the chapter reviews seven significant applied ENLP models, assessing how each fairs in relation to this specification.

3.2.2 Nenov and Dyer (1988, 1993, 1994)

Nenov and Dyer's DETE — which translates as the word *child* in Bulgarian (Nenov's mother tongue) — is described by the authors as a computational model of associative interactions between two cognitive modalities (visual and verbal), devised to explore how language semantics maps to sensory experiences. The authors first paper (Nenov & Dyer 1988) provides a basic overview of the task domain and the practical objective of the research — learning to associate language descriptions of objects moving in a visual field with those objects (see, for example, figure 3.5). Two later papers (Nenov & Dyer 1993, Nenov & Dyer

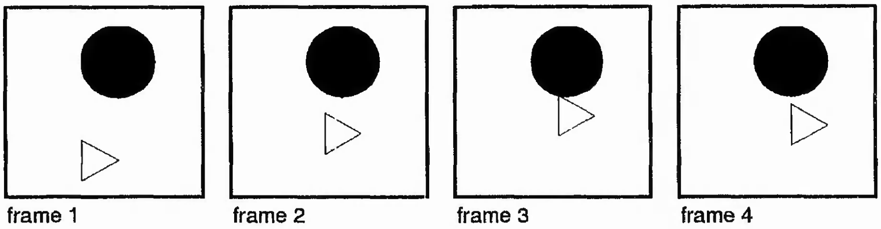

Figure 3.5: An example VS sequence of events to be associated with the verbal sequence 'small triangle moves up, hits black circle and bounces.'

1994) provide detail of the actual implementation of DETE that uses a novel connectionist architecture, termed Katamic Memory, and an account of results arising from a number of investigations with DETE.

With environmental interaction in mind Nenov and Dyer employ a simulated domain called a 'blobs world' which they implement by way of a 64 x 64 pixel visual space (VS). Three blob (object) types (circle, square, triangle) of varying size and colour have the ability to move about the VS in varying directions and at varying speeds. To describe the objects, relations, and events occurring in the VS the authors utilize a fifty word lexicon (Nenov & Dyer 1994, table II.).

DETE is visually interfaced to the world (i.e. the VS) by way of a single circular retina, called the visual field (VF). The visual information captured within the VF is passed to five procedural modules[6] that are attuned to the recognition of shape, size, colour, location, and motion respectively. Each module assesses the presence of certain pre-defined features and produces a binary feature vector, based upon its findings, that is supplied as input to the Katamic memory. These vectors are actually encoded in terms of five feature planes (FPs), each FP being a 2D array of 16 x 16 (256) binary units. An example of representation within a specific FP is cited as follows:

> "In the siZe FP the largest objects (64 pixels on the VS) cause activation in the upper left-hand corner while the smallest (3 pixels) cause activation in the lower right-hand corner" (Nenov & Dyer 1994, p.11)

[6] The authors note that these modules are all specifically designed and explicitly programed (i.e. Designer-Centric) non-connectionist rule based systems.

A further procedural module, called the Word Encoding Mechanism (WEM), provides an interface between DETE and language input. The WEM takes as input typed text and produces as output a sequence of gra-phonemes. There are 26 gra-phonemes, each corresponding to a letter of the English Alphabet, each represented as a 64 bit binary vector. Hence, DETE does not receive complete word representations as input, but rather timed sequences of 64-bit wide distributed patterns. Yet another procedural module, the Verbal Activity Decoder (VAD), works contrary to the WEM producing language output on the basis of gra-phonemic patterns.

The authors adopted an incremental approach to teach DETE the appropriate visual-verbal associations. Initially single word to visual associations were taught starting with the shape words 'circle', 'square' and 'triangle.' VS scenes comprising a single shape, but with varying colours, sizes, locations, and motions, were presented with the appropriate gra-phonemic verbal sequence. Over a number of trials (i.e. different VS scenes) DETE was able to extract the invariant features from amongst the FPs (i.e. the shape) and form an association between these features from the shape FP and the verbal sequence (e.g. 'circle'). Using the same approach, DETE was taught the meanings of words refering to colour, size, location, and motion. Finally, words relating to events, e.g. 'bounces,' were taught (see figure 3.5).

The breadth and scope of the DETE research work initially dazzles — it is without doubt the largest-scale effort in its category to date. However, a thorough and systematic analysis soon leads one to the conclusion that enough has been left vague and unstated, that it is difficult to assess exactly how DETE works as the authors claim that it does. Despite this fact, a number of constructive objections to the approach are now raised based upon the details that are available.

The first objection is directed at the procedural modules and FPs used to process visual input. Initially one can be easily fooled into thinking that DETE is a fully fledged LAG architecture. However, upon a closer inspection one soon realizes that many of the meanings within the system are to some extent specified a priori (cf. (Dorffner 1994)). Consider for example the shape words 'circle', 'square' and 'triangle.' DETE does not *learn* what each shape is in terms of invariant VF input (i.e. it does not categorize sensory data). Instead a procedural module determines, according to some explicit system of rules, whether VF input matches one of three *pre-defined* shape templates. This approach compromises lexical development, since DETE will be unable to learn the meaning of a shape word such as 'rectangle' unless Nenov and Dyer explicitly add a new rectangle shape template and rewrite the input-output rules of the appropriate procedural module. DETE cannot develop its semantic knowledge autonomously as it comes into contact with new VS phenomena.

Another concern — and perhaps the reason as to why DETE cannot directly categorize sensory input — can be levelled at Katamic memory's ability to handle 1-bit densities, i.e. the number of on bits (value=1) within a binary input vector. It would seem from the authors' under-specified results that Katamic memory runs into problems when 1-bit densities exceed 40% (Nenov & Dyer 1993, p.129). If this is indeed the case, then Katamic memory would not be able to deal with realistic sensory data.

One final objection arises due to the complexity of the model's neurobiologically inspired Katamic memory. The results show that in terms of computation, learning is extremely intensive, e.g. it takes DETE about an hour to learn the appropriate associations for three separate words on a CM-2 parallel computer that has 16,000 physical processors, and over a million virtual processors (Nenov & Dyer 1994, p.37). Presently, computational power of CM-2 magnitude is, in most scenarios prohibitively expensive, and as a consequence, one may argue against the practical applicability of the approach.

Criticism aside, there are many positive aspects to the model. Firstly, DETE builds two way associations that allow it to perform both verbal-to-visual association (identification) and visual-to-verbal association (description). Secondly, Katamic memory has temporal qualities that enable it to learn pattern sequences. This ability was put to good use in relation to the word 'bounces.' Unlike for many of the other words there was no representation for the meaning of 'bounces' built into the system priori. Its meaning was acquired through the extraction of an invariant sequence of VS events : (1) blob moving in some direction D_1, folowed by (2) blob not moving, followed by (3) blob moving in different direction D_2. Thirdly, Katamic memory has a built in mechanism that allows it to switch between learning and prediction on a pattern-by-pattern basis. This offers a distinct advantage over other connectionist architectures that require a long learning phase followed by a distinct performance phase (this issue is addressed further in the next chapter). Finally, DETE is able to ground lexical symbols with respect to a number of input modalities, e.g. colour, shape, size, etc. This multi-modal aspect of DETE is discussed further in chapter 7.

3.2.3 Bartell and Cottrell (1991)

The work of Bartell and Cottrell (Bartell & Cottrell 1991) has similarities to DETE in that it too explores associations between visual and verbal input. However, the authors visual domain, a 'billiard ball world', is quite restricted in comparison to the 'blobs world' and as a consequence the vocabulary required to describe events within the world is greatly reduced. The billiard ball world is an environment in which a single billiard ball may roll around a square table and bounce off the table's walled edges. The actual objective of the work is to learn desriptions of movies based upon this environment, where each movie consists of 20 snap-shots of the ball in successive positions on the table.

To address their objective, the authors present a structured connectionist architecture (see figure 3.6), called the Movie Description Network (MDN), which is an amalgamation of two Simple Recurrent Networks (SRNs) (Elman 1990). The application of the architecture to the task was as follows. The input to the MDN was a 2D vector that specified the billiard ball's x, and y position on the table[7] for a particular snap-shot. The output layer consisted of six nodes, each tuned to a single word in the set $\{$*up, down, left, right, slowly, quickly*$\}$.

The MDN was trained, using a supervised back-propagation algorithm, for seven iterations through 50,000 randomly generated movies, i.e. movies in which the starting position $< x, y >$ of the ball was chosen randomly, and the starting velocities δx and $\delta y >$ were chosen randomly in the range [-0.3, +0.3]. Training followed two stages: first the image SRN was trained to predict the next visual state i.e. given $< x_t, y_t >$ predict $< x_{t+1}, y_{t+1} >$; then at random intervals, the word SRN was trained with the appropriate words relating to the sequence of snap-shots sent to the word SRN buffer layer.

To review performance the MDN was tested on a further 1,000 randomly generated movies. 250 of these movies were 'extended' in the sense that they contained 50 snap-shots. A further 500 were 'extended' inasmuch as the parameters δx, and δy set outside the [-0.3, +0.3] range used in training. The reason for using these movies was to test how well the MDN had conceptualized the training data, i.e. acquired the invariant aspects of the input. If the architecture had acquired accurate concepts (meanings), then it would be able to

[7] Values for x, and y lay within the bounds (i.e. walls) of the table, which were positioned at [-1.0, +1.0] in each dimension.

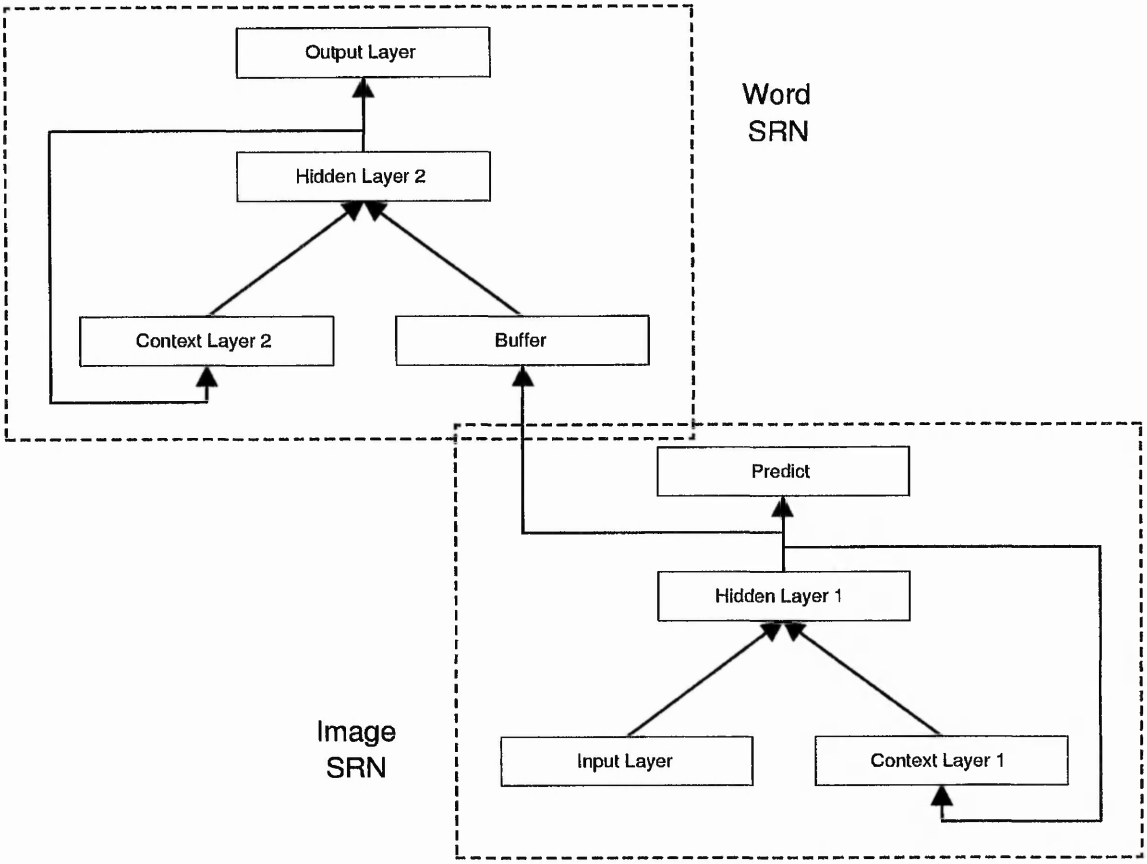

Figure 3.6: Architecture of the Movie Description Network (MDN). A two node input layer processes features x, and y taken from a movie at time, t. The image SRN is taught to predict the value of these features at time, $t+1$. The word SRN then learns the required relations between the time based representations acquired at hidden layer 1 (via buffer layer) and six ouput nodes, each allied to a lexical symbol.

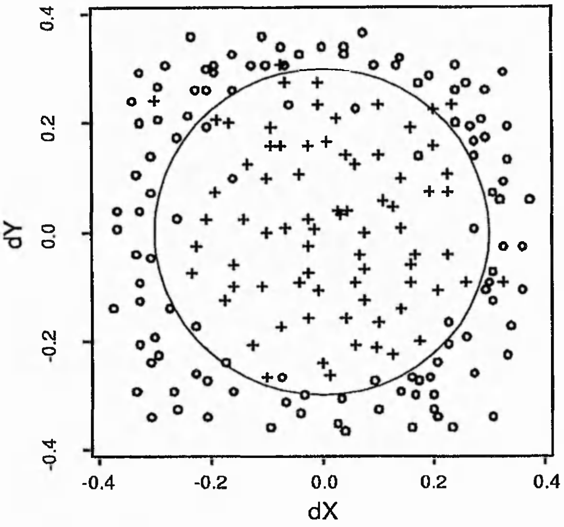

Figure 3.7: After training was complete Bartell and Cottrell analysed the knowledge acquired by the MDN using various techniques (Bartell & Cottrell 1991, p.808). The authors present a figure similar to that shown above that depicts results from the analysis. It shows the internal representational boundary acquired with respect to the lexical symbols 'slowly' (points inside the boundary) and 'quickly' (outside the boundary).

accurately describe an extended movies content. Based upon their testing (i.e. running the CM in its prediction mode), the authors cite quite satisfactory results. For over 80% of the 1,000 test movies the MDN produced an entirely accurate description more than 85% of the time.

What is particularly interesting about the model is that the concepts that it accurately acquires are not based directly upon the input. The MDN's temporal nature (see the delay routes in figure 3.6), enables it to learn concepts that are time dependent. Hence, in addition to creating boundaries in the feature space (x, y), the MDN is able to acquire boundaries in the space $(\delta x, \delta y)$. Figure 3.7 shows how a decision boundary separates the concepts *quickly* and *slowly*. Whilst temporal processing is an attribute for which one should give the MDN model credit, this is more than offset in relation to the following objections:

- The MDN architecture is only able to perform description. Strictly feed-forward processing inhibits its ability to perform identification when given lexical input alone.
- The lexical elements within the model are not acquired. They are set explicitly by the authors in relation to nodes in the output layer.

- The lexicon acquired by the MDN was far too limited consisting of only six words. Furthermore, the quantity of training data required in order to learn the meanings for these words seems excessive: (50,000 movies x 20 snap-shots x 7 epochs) = 7,000,000 input vectors (Bartell & Cottrell 1991, p.807).

- The final objections concern the use of SRN's: a hybrid of the popular MLP CM. The gradient descent supervised learning algorithms (e.g. back-propagation (Rumelhart et al. 1986)) typically used in conjunction with these MLP type achitectures attempt to minimize error by updating the majority, if not all, of the weights in the system in a single learning step. This *fully-distributed* learning — and consequently, knowledge representation — typically results in (1) extremely lengthy training times that can easily run to the order of days (cf. (Sales 1996)), and (2) developmental deficiencies: once such a CM has been trained, i.e. acquired certain concepts, symbols, etc. attempting to teach more to the CM will usually result the corruption of previously learnt knowledge. One might alternatively say that in such models the ability to incrementally acquire and adapt knowledge is seriously compromised. These final issues are discussed at much greater length in the following chapter.

3.2.4 Dorffner (1992)

Following in the footsteps of DETE, and the MDN, Dorffner's Sub-symbolic Language Model (SLM) (Dorffner 1992) also addresses visual to verbal association. However, Dorffner investigates lexical acquisition and grounding in relation to a static visual domain, i.e. individual images, rather than a dynamic movie type visual domain.

The SLM connectionist architecture, shown in figure 3.8, provides what is perhaps the best working example of the generic LAG architecture introduced above. Environmental stimuli are captured and encoded at the SLMs sensory layers. This S-layer activity is then propagated to a C-layer of units. Because all of the weights between these layers, and hence the boundaries, are randomly initialized, the S-layer activity causes a random fuzzy vector of activation across the C-layer units (see figure 3.9). Intra-layer competition is then employed which 'focuses' the fuzzy vector into what is termed an identifiable state. Finally, an unsupervised learning process adapts each weight in proportion to the activity of the

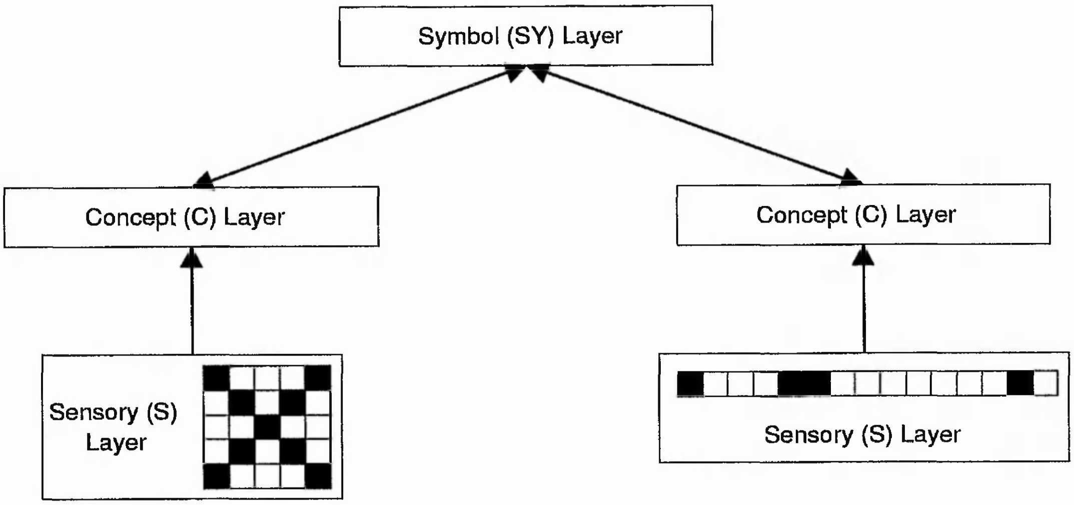

Figure 3.8: Dorffner's sub-symbolic language model employs five layers of processing units in total. Two input, or sensory (S) layers, pass visual and pseudo-acoustic (verbal) input respectivly to two conceptual (C) layers. A Symbol (SY) layer connects the C-layers providing the necessary structure to enable visual concepts to be linked to verbal concepts.

S-unit and C-unit that the weight connects. This competition and learning enable C-layer units to become tuned to the salient, or invariant properties, inherent within a group of sensory patterns, e.g. a sub-pattern that appears within a group of distinct visual images.

Another layer of connectionist computing units employed in the SLM is the symbol (SY) layer. Its chief function is to acquire mappings between units in the C-layers to which it is connected. Strong mappings are acquired between C-units that consistently co-occur in relation to each other, or in other words lexical concepts are bound to visual concepts. Each unit within the SY-layer is responsible for one such mapping.

One interesting aspect of natural language investigated in connection with the SLM is lexical taxonomy. As the author notes: "Many categories that have labels in a language are believed to exist on a hierarchy of subordinate and superordinate categories. Categories on a higher level of this hierarchy (taxonomy) are said to contain all categories on lower levels. Thus, any given object can be named with several labels from different levels of the taxonomy" (Dorffner 1992, p.304). For example, a certain dog might be described as a 'poodle' (subordinate), a 'dog' (basic level), or an 'animal' (superordinate). Dorffner provides a simple example of the SLMs functionality. The model is applied to the task

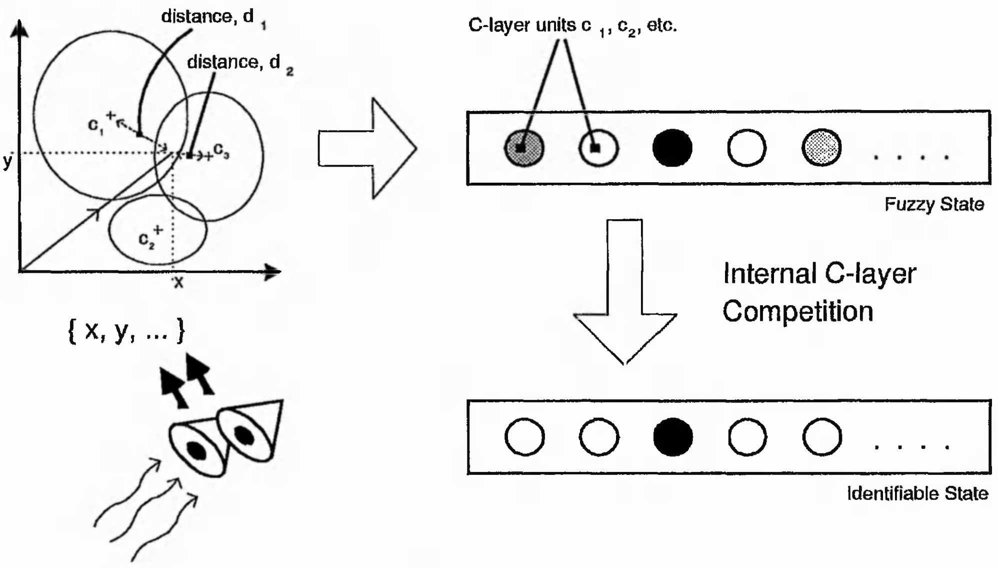

Figure 3.9: Show (1) an image (2) how this relates to a point in vector space (in reality this will be a 25D space, however, for demonstative purposes only 2D shown) (3) put a number of circular boundaries in vector space, (4) show how they relate to the C-layer units in a fuzzy-state. (a) a fuzzy state (b) an identifiable state

of acquiring and grounding seven lexical items : two basic level – 'table,' 'chair;' four subordinate – 'baroque (table),' 'kitchen (table),' 'arm (chair),' 'garden (chair);' and one superordinate – 'furniture.' Each is presented to the right-hand side of the SLM in a vector encoded format. The visual input consists of noisy[8] versions of four 5x5 pixel images. Each image was presented in turn with each of the lexical items that correctly described it.

Further details of the investigation and the results obtained are sketchy to say the least. The author provides nothing in the way of empirical facts and figures. However, on the basis of what there is, a number of objections are raised below. Firstly, the author briefly notes that the SLM failed to aquire a few of the required mappings [p.308] between concepts, and furthermore that other mappings were too weak. In such a simple and limited domain of investigation (i.e. 7 words, 4 images) such results are a cause for concern.

The author also mentions on a number of occasions that for a given C-layer a certain grouping of patterns (a concept) will only be discovered with a certain probability. To

[8] Up to 20% noise was added to the input images. This was to show that the SLM could generalize and conceptualize as required.

combat this problem the use of multiple C-layers connected to each S-layer is suggested. However, in Dorffner's analysis there is no suggestion as to the number of layers that might be required for a given task, i.e. there is a lack of design methodology. In his own example investigation the SLM employed five C-layers, and 4 SY-layers, requiring a total of 3870 connections (weights). But given the failures, Dorffner suggests more layers *may* be required. As a related point, given the very limited nature of the task at hand in relation to such an extensive architecture, one must question the computational efficiency of Dorffner's approach.

A futher issue that arises briefly in Dorffner's analysis is that of word frequency. The fact that the verbal input 'furniture' was presented to the SLM on more occasions than some of the other verbal inputs seemed to cause some problems. In a real world application any 'balanced frequency' constraint would be unacceptable.

Finally, as with the SRN used in the previously described model, the fully-distributed nature of the SLM effectivly restricts the developmental capacity of the model. Furthermore, other research has shown that the Interactive Activation algorithm (McClelland & Rumelhart 1981), employed to effect C-layer competition, is flawed in its ability to maintain stable representations (Grossberg 1987).

On the positive side, the SLM is able to trigger verbal concepts given visual input alone (description) and vice versa (identification). Another favourable aspect of the model is its conformance to autonomous principles, i.e. symbols etc. are not explicitly built into the model a priori. The final quality worthy of mention is the architectures ability to deal directly with raw unprocessed sensory stimuli, e.g. CCD camera ouput. Such input was not used in conjunction with DETE or the MDN.

3.2.5 Plunkett et al. (1992)

Plunkett, Sinha, Strandsby and Møller (Plunkett et al. 1992) cite their work as an investigation of concept formation and vocabulary growth. The domain of this investigation centers, as before, on the problem of associating static visual input and linguistic input. The work is actually based upon the work of Chauvin (Chauvin 1989), however, the authors note that whilst the architecture of the CM utilized and the nature of the problem tackled are similar, there are crucial differences in terms of the scale of the learning problem (Chauvin worked with a mere 4 lexical symbols) and the manner in which the performance of the model is evaluated.

Figure 3.10a depicts the structure of the CM employed by the authors — one can again see some resemblance to the generic LAG architecture introduced above — to tackle the specific task of associating visual input from a 19x9 retinal array, and lexical input, taking the form of 32 bit binary vectors. The visual image data was generated from 32 prototype images. Six noisy versions of each image constituted the 192 visual patterns used in the investigation. The lexical input data comprised 32 orthogonal vectors (i.e. 32 lexical symbols), each having a single bit active. Training the CM to complete the task consisted of three separate stages[9]:

1. The left half of the CM was trained (see figure 3.10b). A visual pattern is presented across the retinal input units and activity is propagated through the CMs layers to the retinal output units. The activity across the output units is compared to the input unit activity. The difference between the activities is recorded as an error. The supervised back-propagation algorithm was used to adapt the weights on the visual side of the CM so as to reduce the error. Hence, the CM was trained to reproduce a visual input at the output — an auto-associative learning task.
2. Similarly, the lexically oriented right half of the CM was trained (see figure 3.10c).
3. As a final step the CM was trained as a whole. Visual inputs and their corresponding lexical inputs are presented simultaneously. Errors at the output units are minimised by adapting all of the CMs weights.

[9] An in-depth investigation of the three stage training referred to as CP training can be found in (Sales 1996). A number of benefits in relation to a single step training strategy are noted.

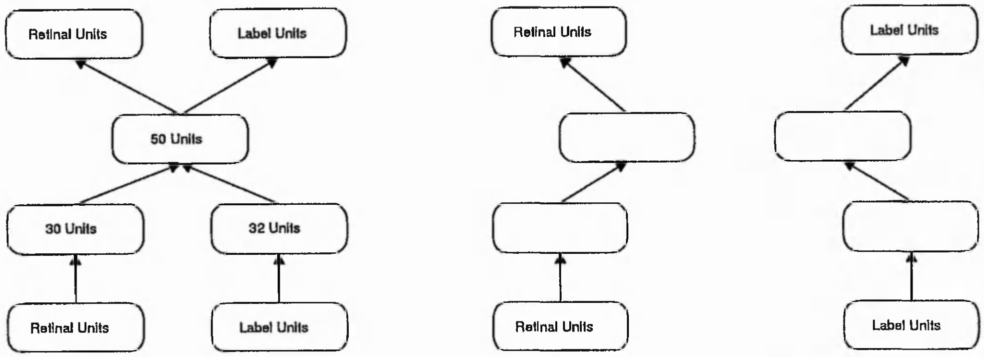

Figure 3.10: (a) The seven layer connectionist architecture employed by Plunkett et al. Input from each sensory modality is propagated to a *distinct* layer of hidden units. The output from both layers then converges to a second *shared* layer of units. (b) The first (visual) phase of auto-associative training (refer to main text for a full description). (c) The second (lexical) phase of auto-associative training. Figure reproduced from (Plunkett et al. 1992, p.300).

The authors' analysis of the model is based upon two measures of performance : 'production,' and 'comprehension,' which correspond directly, and respectively to the functions of *description* and *identification.* After each 3-phase cycle of training (one epoch) the production and comprehension performance of the CM was assessed. All in all the model was trained for a total of 100 epochs. The 100 comprehension and production scores were presented in a graphical form.

Whilst the authors' own interests lie with the characteristics of the comprehension and production graphs, in particular how they tie in with the characteristics of vocabulary growth exhibited by young children with developing linguistic and conceptual systems, here one is more concerned with the actual figures that were achieved. The actual results show that the best performance values were recorded when the CM was fully trained (i.e. after 100 epochs). At this point production and comprehension accuracies of around 85% and 83% were achieved in relation to the original prototype images. Given that the un-noisy images were never presented to the CM during training, the results above lend themselves to the claim that the model has generalized correctly on the basis of input stimuli and acquired appropriate concepts, i.e. the model has correctly partitioned its feature space.

The main criticisms of the work centres around the author's choice of connectionist model. In common with the SRNs employed by Bartell and Cottrell, the CM employed by Plunkett et al. is an MLP hybrid trained by back-propagation learning (Rumelhart et al. 1986). As mentioned above, such an architecture/algorithm combination brings problems in terms of slow training times and the ability to incrementally acquire and adapt concepts. In addition to these problems there is the issue of design methodology. There are no hard and fast rules when it comes to designing and applying MLP type models — the topology of the architecture, the number of nodes in each layer, the initial values of weights and other learning algorithm parameters, all depend upon the task at hand (i.e. the data) and are all typically determined through trial and error empirical exploration. Although Plunkett et al. make no mention of their particular route of exploration, Sales (Sales 1996) has shown that production and comprehension results vary widely given different parameterizations. As a consequence, the design process can be, and often is, an arduous and lengthy one.

One further point of interest in relation to this work is the author's observation that in their model there is a linguistic influence on non-linguistic conceptualization:

> "... the network exploits the predictive power of the input label in identifying category membership, as well as the natural clustering of the distortions in the image plane" (Plunkett et al. 1992, p.305).

This is a contentious issue in linguistics, cognitive science etc. The arguments for and against are considered further in chapter 5 as a basis from which to discuss the engineering impact that such influence can have *vis-à-vis* the design of artificial ENLP systems.

3.2.6 Lammens (1994)

Lammens' (Lammens 1994) model of colour perception and naming (MCPN), is the only piece of applied work reviewed in this chapter that is not built upon connectionist foundations. However, the underlying substrate of computation and representation is spatial and as such closely resembles that of a CM. Lammens' work focuses on the definition of a referential semantic model of (basic) colour terms, or from a mathematical perspective on the construction of a mapping $\aleph$,

$$\aleph : \mathbf{E} \mapsto \mathbf{C} \times \mathbf{I}$$

that connects a set of possible visual stimuli, **E** to a set of eleven colour terms (lexical symbols), **C**:

$$C = \{white, black, red, green, yellow, blue, brown, purple, pink, orange, gray\}$$

where **I** represents the closed interval [0, 1] that allows one to define a 'goodness' rating between a visual stimulus $E(\lambda) \in \mathbf{E}$ and a particular colour symbol $C \in \mathbf{C}$. Hence, the domain of investigation is again one of visual/linguistic association.

Lammens' work is important and interesting for its further analysis of embodied semantic systems. Given an embodied view of semantics where one forfeits the notion of objective meaning in favour of internal subjective meaning, it is vital that communicating bodies have (a) isomorphic symbol to concept groundings and (b) roughly isomorphic concepts, i.e. language is of little use if person A refers to concept X using the word W if person B interprets W as relating to concept X' (cf. (Winograd 1980)). Lammens argues that the organization of colour experience (i.e., into colour concepts) within humans is determined largely by the structure of human colour receptors, and that since this structure is by and large standard for all members of the species, these colour concepts are more or less isomorphic. As a consequence, the author notes that "To define an adequate [artificial] model of the semantics of colour terms in natural languages, it is necessary to model the physiology of human colour perception" (Lammens 1994, p.16).

Constructing $\aleph$ is specified as a 2 part task. The first part of the mapping deals with the transformation of a visual stimulus into a point in a colour space (i.e. a conceptual

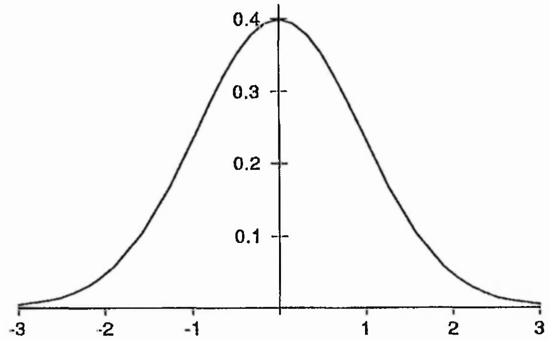

Figure 3.11: The above plot shows a 1-D Gauusian function. μ places the centre of the category, σ defines the spread.

space that organizes colour experience). Based upon various neurophysiological findings the author proposes a three dimensional colour space (referred to as the NPP space) and a function that projects a physical stimulus $E(\lambda)$ — electro-magnetic radiation in the visible wavelength range [$\approx$ 380-770 nm] — onto a point $\overline{p}$ in the NPP colour space.

The second part of the mapping concerns the creation of categorical regions defined within the colour space, each paired with a unique colour symbol. Put simply, it deals with a move from a colour space to colour names. Using data acquired by Berlin and Kay (Berlin & Kay 1969) on how human subjects name colours, the author fitted eleven categories within the NPP space. Each category was based upon a normalized Gaussian function, and was thus modulated by way of two parameters: μ, and σ (see figure 3.11). The trial-and-error fitting process involved trying different values of these parameters in order to minimise a specific error criterion (Lammens 1994, p.129).

Lammens reserved some of the Berlin and Kay data in order to test how well the referential semantic model could name unseen colour stimuli, i.e. perform description. Given a novel stimuli and applying the appropriate transformation function one can determine its position $\overline{p}$ in NPP space. Given $\overline{p}$ one can then determine:

$$\alpha_i = G_n(\overline{p}, \overline{\mu_i}, \sigma_i) \tag{3.1}$$

where α_i provides the likelihood that the stimuli belongs to category number i, G_n is a normalised Gaussian function. The highest of the eleven α values was selected, and provided

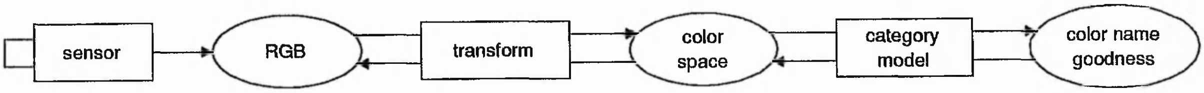

Figure 3.12: Lammens' colour naming process. Boxes represent devices or operations, ellipse represent data representations (see text for details). Figure reproduced from (Lammens 1994, p.146).

that it exceeded a threshold θ, the category C_i affiliated with the value was chosen. Finally, the lexical symbol paired with C_i is elected to name the stimuli. Figure 3.12 provides a schematic depiction of the colour naming process.

The author presents a number of quantitative results that seem to provide evidence of success; in particular the error rates for blacks, greys, and oranges are almost zero. It should be noted, however, that the author makes no reference as to the number of test stimuli used in testing, nor to the distribution of these stimuli, and as such one could question the relevance of the results.

The most important qualities of the MCPN are (1) its ability to perform description and identification — the model could also be used in reverse, i.e. to point out examples of colours — and (2) the topological nature of the NPP colour space. The NPP space, like other physiologically and psychologically inspired colour spaces (more on these in chapter 5), directly captures a certain order between different colours that humans introspectively perceive. For example, the Gaussian enclosed cluster of points associated with the label 'red' is closer to the cluster associated with 'orange' than it is to the one associated with 'green.' Therefore, one gets a continuity between categories, i.e. there are smooth transitions from one category to another as one moves linearly through the space from point $\overline{p}$ to $\overline{p'}$. As a result of this the MCPN was able to produce appropriate complex names for 'borderline' colours, e.g. where two α values exceed θ:

$$\alpha_4 = 0.75 \tag{3.2}$$

$$\alpha_6 = 0.92 \tag{3.3}$$

the model produced a compound label 'greenish blue.'

Furthermore, one can argue that such a space also implicitly encodes meaning postulates (see section 2.2.1) without any designer input, e.g. there is no need for a rule that explicitly states that if something is red then it cannot be green ($\forall x[red(x) \not\rightarrow green(x)]$). The very fact that the green and red categories are not neighbours within the NPP space, means that the label 'greenish red' will be considered as semantically strange. In chapter 5 further issues relating to the topology of conceptual spaces are considered, in particular, how the structure of the space aids the learning of concepts within a connectionist framework.

The main negative element of the work is of course its neglect of development. The semantic knowledge in the model is not acquired, but is provided explicitely by the designer, hence, the model is one of lexical grounding rather than lexical acquisition and grounding. The model is thus constrained in its abilities. It would be impossible for the MCPN to develop linguistically as its experience of a domain develops, e.g. in much the same way that an artist acquires an extended colour vocabulary and talks of Ochres and Sienna.

3.2.7 Grumbach (1996)

Grumbach (Grumbach 1996) describes a hybrid model of symbol formation whose task domain focuses around a robot learning situation. Specifically, the objective of the work was to assess how grounded language terms could aid in training a mobile robot to follow a road whilst avoiding obstacles. A simulation was constructed of the domain, which contained a mobile robot which was given visual sensors and the ability to perform one of seven actions; increase speed, decrease speed, no change, speed up and turn left, speed up and turn right, slow down and turn left, slow down and turn right.

Grumbach proposed a two stage training regime for his mobile robot. The first was qualified as a *learning by example* (LBE) stage and consisted of training 13 typical static visual inputs the robot may encounter to produce one of the seven possible actions. The LBE stage was implemented using a simple associative CM with only two layers; a (visual) input layer and an (action) output layer consisting of seven units. Subsequent to LBE the robot was tested on two roads. On a simple circular road, performance was found to be satisfactory. However, on a more intricate road containing a number of sharp bends, the robot failed to keep to the desired path on a number of occasions. In an attempt to correct these failures

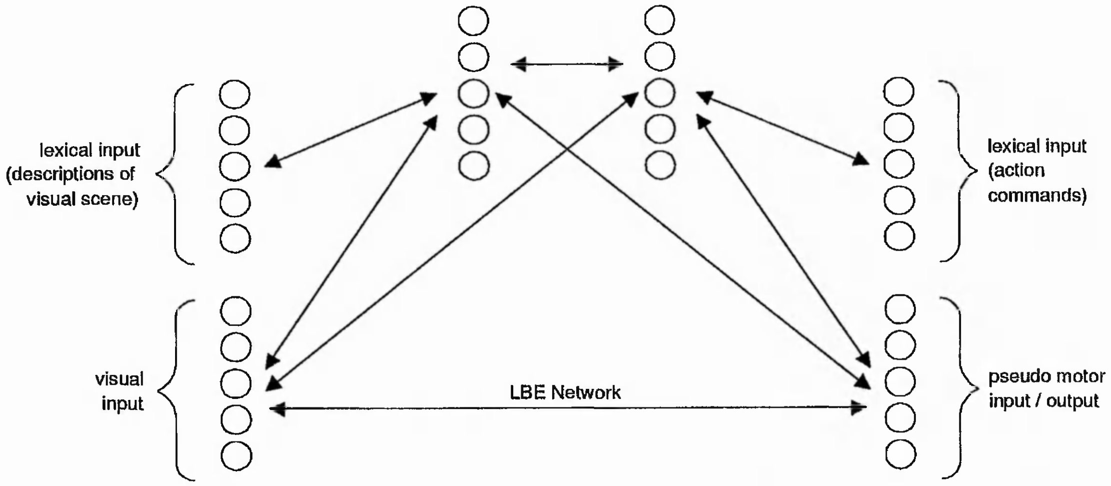

Figure 3.13: The learning architecture shows the parallel routes between the visual input and the action output, the first via the initially trained LBE network, the others via the LBI networks. LBE net trained by delta-rule, LBIs by simple Hebbian learning.

the authors proposed a second stage of training referred to as *learning by instruction* (LBI), where directions could be given to the robot using natural-language, hence, the requirement for a set of grounded lexical symbols.

Figure 3.13 depicts the model employed by Grumbach which consists of a number of simple associative CMs. Preceding LBI symbol acquisition and grounding was initiated requiring two of the simple CMs. The first was trained to associate the visual input from the robot with a verbal label describing it, e.g. *left bend* whilst the second was trained to associate each of the seven possible actions with a verbal description, e.g. *slow down*. With the grounded symbols acquired, another simple CM was then utilized to implement LBI by forming connections between the two sets of grounded symbols, i.e. the visual labels and the action labels. The appropriate connections were trained by supplying pairs of lexical items together, e.g. training the instruction "if the road bends sharply to the left, you must slow down and turn to the left", required that the lexical items *sharp bend* and *slow down and turn left* be presented to the CM simultaneously.

Grumbach's model is lacking from an embodied semantic perspective on a number of fronts. The main weakness stems from the lack of conceptual content established within the model. The simple CMs employed by the author are too simple, in fact, given the overview above one could argue that they do not even qualify as CMs. These linear associative devices that utilize basic Hebbian learning simply connect sensory representations directly to arbitrary symbolic units. They possess no conceptual space as such, nor do they optimize categorical regions within this space. This means that the system must explicitly learn the mappings between a symbol and every possible sensory percept to which it may apply. The whole incentive for using CMs as a basis for embodied semantic systems (i.e., their ability to categorize, generalize etc.), seems lost on Grumbach. In his introductory preamble Grumbach speaks of the model acquiring concepts [p.133], however, given the above analysis this is clearly not the case.

One aspect of Grumbach's model that initially looks promising, is its ability to ground symbols with respect to more than one input modality, i.e. verbal to visual association, *and* verbal to action association. However, this multi-modal aspect of the model turns out to be an Achilles' heel. The author uses two separate processing channels for verbal input, one for words relating to visual input, the other for words relating to actions. Consequently, one may argue that the author is making a priori assumptions regarding the meanings of the lexical items. This negative aspect of the model is discussed further in chapter 7

3.2.8 Sales (1996)

The domain of investigation proposed by Sales (Sales 1996) is an adaptation of that examined by Plunkett et al. — the core objective being the creation of a system that can acquire a grounded lexicon autonomously through embodied interaction in a verbal and visual environment. However, Sales' research stands in contrast to the research that has been discussed above in that it is 'application-oriented'. In this sense LAG is not addressed purely as an academic task. Rather, the author is interested in how an ENLP system can be applied in a practical setting, e.g. as part of an intelligent navigation system for a robot. As such special attention is paid to issues which affect practical applicability.

The practical approach adopted by Sales is very similar in principle to that ventured in this thesis (see section 3.2.1). Since the aims of both pieces of work are quite similar, a detailed analysis of Sales' models, investigations and results is deferred until chapter 6 where they are considered as a base for comparative study. What follows below is a brief overview.

Sales proposes a lexical acquisition and grounding model based upon *Weight-less Artificial Neural Network* (WANN) technology. WANNs are often not regarded as CMs, mainly because they do not embody the two general characteristics that usually define such models (see section 3.1.1 above). Specifically, WANNs lack both weights and a fan-in function, and as a consequence their operating dynamics, i.e. their learning and prediction functions, are quite different from those described above. However, at an alternative level of interpretation the similarities between WANNs and conventional CMs are plain to see: they are pattern classifying devices, they operate upon n-dimensional vector input, and they adjudge pattern similarities (i.e., generalize) by way of spatial reasoning.

Sales' justifies the choice of a WANN based LAG architecture by way of the following features:

- **Speed Of Acquisition**: Sales suggests that as a result of their alternative dynamics WANNs can acquire knowledge, e.g. lexical and sensory concepts, much faster than many other connectionist systems.
- **Re-learning**: Another positive aspect of WANNs highlighted by the author is their ability to add to their knowledge in an incremental manner. Hence, such systems can gradually develop a meaningful lexicon adding new words as required.

The data used in Sales' investigations did not match that employed by Plunkett et al.This was mainly because WANNs are unable to process analog input, all input must be binary. Instead in new investigations Sales generated his own visual and lexical data: 37 12x12 (144 bit) visual patterns and 37 64 bit lexical patterns. Through the course of extensive investigations Sales employed a total of ninety WANN models. Each model had different connectivity patterns between input and output units, and each was trained using the 37 pattern pair training set and then tested on data obtained by adding 10% noise to the original data.

As above, analysis of the models was based upon *description* and *identification*. In contrast to Plunkett et als. model it was found that performance on these tasks was far from balanced. The best description and identification accuracies recorded were 87.4% and 58.4% by models 3 and 44 respectively. Model 44 also came out top when averaged over the two tasks, recording an identification accuracy of 57.7%.

Starting the the positive aspects of Sales' WANN architectures. Firstly, Sales can rightly boast that in comparison to other models described above WANNs are able to acquire knowledge at a greatly improved rate – it took 30-40 seconds for a WANN to acquire the knowledge it would take about 8 hours for an MLP based model to learn. A second plus point is that WANN model's can acquire knowledge incrementally. However, it should be noted that each WANN has a 'storage limit', and that "performances will begin to degrade as the storage limit is approached" (Sales 1996, p.178).

On the negative side one might raise the following objections:

1. A lack of autonomy is evident in the WANN approach. The results from experimentation show wide performance variations on the same data given varying network topologies. Only trial and error seems to be offered by the author as a methodology for finding which topology works best with a specific data set[10].
2. The WANN approach that Sales offers is only capable of dealing with binary data.
3. WANNs suffer from slow recall of knowledge. In all cases it took WANN models at least three and a half seconds to undertake either description or identification. By comparison an MLP type model can recall knowledge in the order of milliseconds.

3.2.9 Discussion

A total of seven models have been described and discussed in the preceding sections of this chapter. In each case at least one of the requirements detailed in section 3.2.1 was addressed. However, in none of the models were all of the requirements met. Thus one can immediately see that scope for improvement remains in the field of embodied natural language processing models.

[10] This inference is based upon numerous remarks made in Sales' text, e.g. "a certain amount of empirical adjustment will be necessary" (Sales 1996, p.223), "connectivity levels for optimum performance vary with data format" (Sales 1996, p.223)

Chapter 4

Adaptive Resonance Theory: A New Framework for Lexical Acquisition and Grounding

One of the primary roles of the previous chapter was to remark upon the representational affinity that exists between embodied semantic theories and connectionist models of computation, and, as a consequence, to justify in general terms the application of the latter as a platform for the practical implementation of the former. However, despite this general affinity, the latter sections of the chapter clearly indicated that all CMs are not equally suited as implementational tools. Beyond their generic facade one finds that different kinds of CM exhibit and possess many different computational characteristics and constraints. As a result some CMs better suit the development of practical ENLP systems (see requirements, sec. 3.2.1) than others.

This chapter introduces and positions an alternative medium of connectionist computation known as Adaptive Resonance Theory (ART), as a superior means through which to implement an embodied semantic theory and consequentially as a superior means through which to acquire and ground lexical symbols. The sections that follow address ART in some detail in order to justify such a bold claim.

4.1 ART: A Brief Introduction

Adaptive Resonance Theory, or ART, was first introduced in 1976 (Grossberg 1976*a*, Grossberg 1976*b*) and originated from an analysis of both artificial and human cognitive information processing. Starting from a number of core principles, Grossberg, Carpenter, and other colleagues have steadily developed and incorporated new ideas into ART (see, for example, (Carpenter & Grossberg 1994, Williamson 1995)). Furthermore, they have realized these principles and ideas through an ever evolving family of connectionist architectures and learning algorithms that have demonstrated increasingly powerful category learning, recognition, and prediction capabilities.

4.1.1 Stable Learning throughout Life

One of the wonders of human cognitive information processing is its *plasticity*. Human learning is, for the most part, sequential, i.e. one does not learn everything about the world in one shot and then simply recall this knowledge, one acquires knowledge in a step-by-step piecemeal fashion. For example, as an infant one learns to recognize one's parent's faces, a little later in life when starting school one must learn to recognize many new faces (i.e. friends), and still later, when starting one's first job, yet more faces must be learnt. Hence, human learning can be said to be plastic in that it is a life-long process — information acquisition is incremental. Moreover, plasticity also accounts for the fact that knowledge is continually adaptable, e.g. as one's parents grow older, one adapts one's internal knowledge of their looks in order to facilitate recognition. Learned knowledge tracks environmental change.

A second equally important characteristic of human memory and learning is its *stability*. Stability in this context refers to the fact that the acquisition of new knowledge does not wash away, or overwrite prior learning, e.g. one does not return home after one's first day at school (having learnt about many new faces) unable to recognize one's parents. Stability does not necessarily mean that knowledge is not forgotten, simply that any loss of knowledge is progressive rather than instantaneous.

4.1.2 The Stability-Plasticity Dilemma

Most contemporary theories of connectionist processing do not match up to the human processing model. In fact many CMs exhibit a phenomenon that is widely referred to as 'catastrophic interference' — a term first introduced by McCloskey & Cohen (1989) founded upon their computational analysis of the operational properties of Multi-Layer Perceptrons (MLPs). McCloskey and Cohen found that MLP models have an inherent tendency to abruptly and completely forget previously learned information in the presence of new input (i.e. in a sequential learning environment) — new knowledge 'catastrophically interferes' with old knowledge. The authors termed this 'The Sequential Learning Problem'.

Notwithstanding the fact that McCloskey and Cohen's published work has done much to raise the profile of the above issues, put into the correct historical context, the author's core findings simply echo those made earlier by Stephen Grossberg. As early as 1976 (Grossberg 1976*a*) findings had been published that a certain class of CMs (viz. feedforward competitive learning models) could, under sequential learning conditions, (i.e. in environments requiring plasticity), exhibit unstable learning. The author proceeded to show (Grossberg 1987, Grossberg 1988) that this instability problem was not particular to competitive models; pointing out that MLP models, such as the back propagation model, and auto-associators, both exhibit similar learning deficiencies. In Grossberg's terms what all of these models failed to solve was what he termed the *stability-plasticity dilemma*[1], a predicament later characterised (Carpenter & Grossberg 1987, Carpenter & Grossberg 1988) by way of a series of questions:

1. How can a learning system remain adaptive (plastic) in response to significant input, yet remain stable in response to irrelevant input?

2. How does the system know to switch between its plastic and its stable modes? (*i.e. where is the boundary between significant and irrelevant?*)

[1] Although Grossberg's analysis pre-dates that of McCloskey and Cohen many still erroneously cite the latter's work when referring to the origin of these findings and give little, if any recognition to the formers findings (e.g., (French 1997)). As a consequence one finds that the latter's terminology (i.e., 'catastrophic interference,' 'sequential learning problem') is generally employed in academic literature (e.g., (Sharkey & Sharkey 1995, Grossberg 1997)).

3. How can the system retain previously learned information while continuing to learn new things?

4.1.3 The Foundations of ART

A formal analysis of how the human learning mechanism addresses the questions raised above, led to and provided the foundations for, the introduction of ART. An analysis of interdisciplinary data arising from various sources (e.g. studies in perception, cognition, attention, and neurophysiology), led Grossberg to postulate that stable and plastic learning is regulated by a matching process, whereby bottom-up stimuli from the outside world are matched against 'top-down learned expectations,' i.e. feedback from conceptual memory (see Grossberg (1976*b*) for further theoretical details).

The theoretical ideas embodied by ART led to the ART family of CMs that have been widely employed in many areas of research and practice. From ART1 (Carpenter & Grossberg 1987) through to Gaussian ARTMAP (Williamson 1995) (the latest in a long-line of ART based CMs), all CMs within the ART family share a set of key characteristics and functional dynamics that are quite unique and set them apart from other CMs (see, for example, (Hanson & Burr 1990, fig. 1, p. 472)). As a result ART based CMs not only solve the plasticity-stability dilemma, but also support fast learning, autonomous self-organization, and rapid knowledge recall. In the final section of this chapter these and other properties of ART models are discussed with respect to lexical acquisition and grounding, however, first the following section addresses, in functional terms, how ART processing mechanisms realize these important properties.

4.2 Learning and Prediction in ART Systems

The continual development of ART has led to the creation of two main classes of connectionist architecture: unsupervised, and supervised. In the latter chapters of this thesis models of the supervised type are employed in an embodied semantic capacity. However, as it will be seen below, supervised ART systems — known more commonly as ARTMAP systems — are built upon the foundations of multiple interacting unsupervised ART CMs, with the former directly inheriting many of the latter's important properties. Hence, it seems logical to first discuss unsupervised ART models.

4.2.1 Unsupervised ART systems

The central feature of all *unsupervised* ART systems (henceforth described simply as ART systems/models etc.), is a unique pattern matching process that compares the current input with a learned category representation, or active hypothesis, selected by the input. The matching process leads either directly to a resonant state, which in turn triggers category learning, or to a search cycle that always leads to a resonant state, unless memory capacity is exceeded. Figure 4.1 illustrates the general form of an ART system and figure 4.2 illustrates the mechanics of a typical ART processing cycle. The three possible stages of this cycle are outlined briefly in the text that follows[2].

I. Bottom-up Activation and F_2 Choice

A preprocessed[3] input pattern **I** registers itself as a pattern **X** of STM activity across the units in F_1 (see figs. 4.1 & 4.2a). The F_1 activity is then sent via multiple fan-ins of adaptive bottom-up weights to F_2. Each committed[4] category unit in F_2 aggregates the signals it receives by way of a fan-in function (as described in sec. 3.1.1). This results in a vector of activity **T** across F_2. The competitive dynamics of F_2 contrast-enhance **T**, i.e. competition *chooses* the F_2 unit J that receives maximal $F_1 \rightarrow F_2$ input, where

$$T_J = max\{T_j : j = 1, \ldots, N\}$$

[2] For a much greater insight into ART systems (e.g. their basis in theory; the technical details of implementation) the reader is referred to the following works: Carpenter & Grossberg (1987), Grossberg (1988), Carpenter & Grossberg (1994), Carpenter et al. (1996).

[3] Each input pattern may be the output pattern of a preprocessing stage. Different preprocessing is given, for example, to speech signals and to visual signals before the outcome of such modality-specific preprocessing ever reaches the attentional subsystem. F_0 can thus be thought of as a filtering field that extracts useful information from original stimuli.

[4] A committed unit is one which has already undertaken prior learning, i.e. one that has acquired a categorical representation.

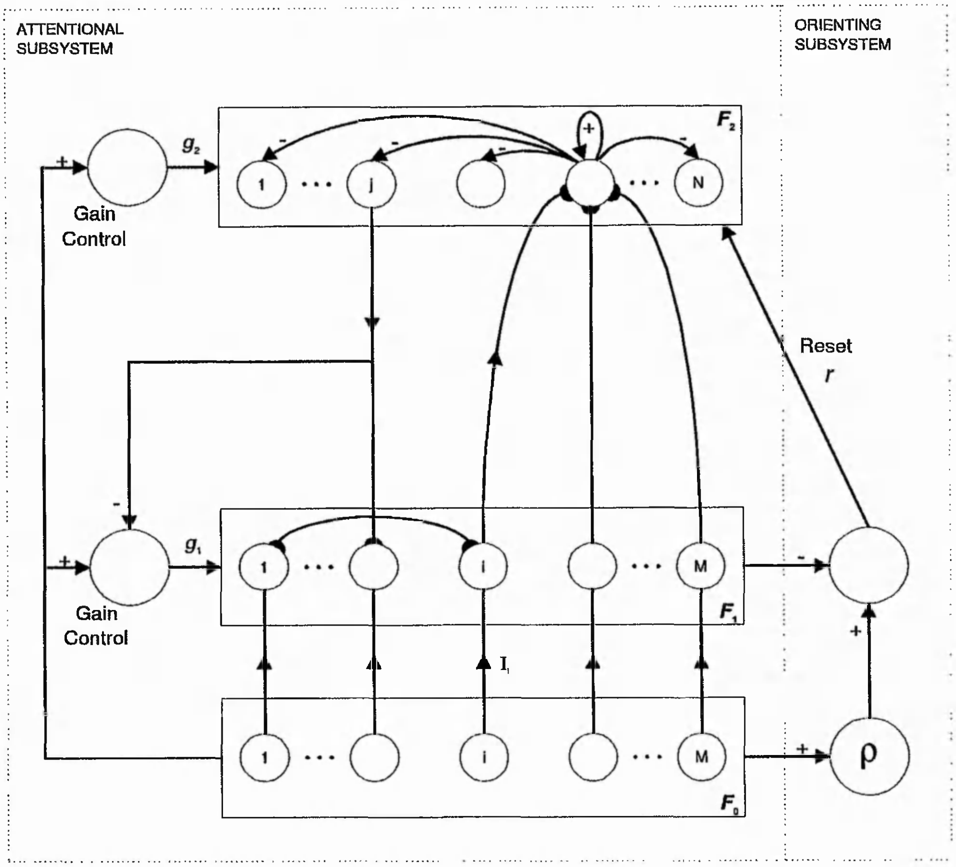

Figure 4.1: Anatomy of an unsupervised ART learning architecture. All such systems possess an attentional subsystem that contains a field (or layer) F_0 of units that represent a current input vector I, a field F_2 that contains category units, and a field F_1 that receives bottom-up input from F_0, and top-down input from F_2. The F_2 field has competitive dynamics – category units compete for activity leading typically to a winner-take-all (WTA) situation. The vectors of activation that form across the fields F_1, and F_2 are described as short-term memory (STM) representations. The STM description accounts for the fact that such representations (1) only typically exist during the presence of the input stimuli I, and (2) may be quickly inhibited by other control signals arising within the system, e.g. the reset signal r. In an ART 1 system a set of bottom-up adaptive weights fan-in connecting all F_1 units to each F_2 unit. A further set of top-down adaptive weights fan-out from each F_2 unit to all F_1 units. These sets of weights constitute the systems long-term memory (LTM). The process of learning imparts useful knowledge into the system in terms of LTM weight vectors. LTM information persists after input offset but remains adaptable. The connections between F_0 and F_1 units are one-to-one and are non-weighted. The orienting subsystem and gain control mechanisms are used to *physically* implement pattern matching, novelty assessment, and category search (see fig. 4.2 text for details). A plus sign indicates an excitatory connection, a minus sign an inhibitory one. All LTM connections are excitatory.

The resulting activity vector **Y** across F_2 has only one non-zero component (e.g. {0,1,0,0} in fig. 4.2b – only the 2^{nd} F_2 unit has activity).

II. Top-down Hypothesis Testing

In ART systems the activation of a single 'winning' F_2 unit is deemed as simply 'making a hypothesis' about which category input **I** belongs to. The next stage of processing is to assess whether **I** is sufficiently similar to the chosen category representation in order to warrant this hypothesis. **Y** is transported via top-down weights to F_1 resulting in an $F_2 \rightarrow F_1$ input vector, **V** (e.g. {0,1,1,0,1} in fig. 4.2b). At this point F_1 is receiving two sources of input: the bottom-up input pattern **I**, and the top-down category representation 5**V**. A matching process at F_1 now ensues that compares **I** and **V**. If **I** and **V** are close enough, by some measure (a matching function, or rule), in relation to an ART system parameter ρ called *vigilance*, then the hypothesis is confirmed (a match) and the system is said to enter a 'resonant state.' If they are not close enough then the hypothesis that **I** belongs to category J is disconfirmed (a mismatch), and the ART system searches its memory (i.e. its F_2 layer) for a better 'hypothesis.'

III. Resonance, or Category Reset and Search

The state of resonance that follows a **V-I** match results in learning, whereby the LTM weights that connect the confirmed F_2 winning unit to all F_1 units, may be adapted, i.e. the selected category's learned representation may be refined to incorporate new information from the current input. Alternatively, in the event of a **V-I** mismatch, a reset signal is sent to the winning F_2 unit, which effectively inhibits or turns off the unit. This inhibition results in the removal of the F_2 activity vector **Y**, which in turn leads to the removal of **V** to F_1 (fig. 4.2c). Without **V**, and thus any mismatch between **I** and **V**, the lone bottom-up input **I** may again reinstate the original activity pattern **X** across F_1 which again is fed via bottom-up weights to F_2 (fig. 4.2d). However, the enduring inhibition means that competition leads to the choice of an alternative unit (a new **Y**), and hence an alternative 'hypothesis.' As before, this new hypothesis is tested. The search process ends in one of three ways: (1) A committed F_2 category unit is selected whose top-down category representation **V**

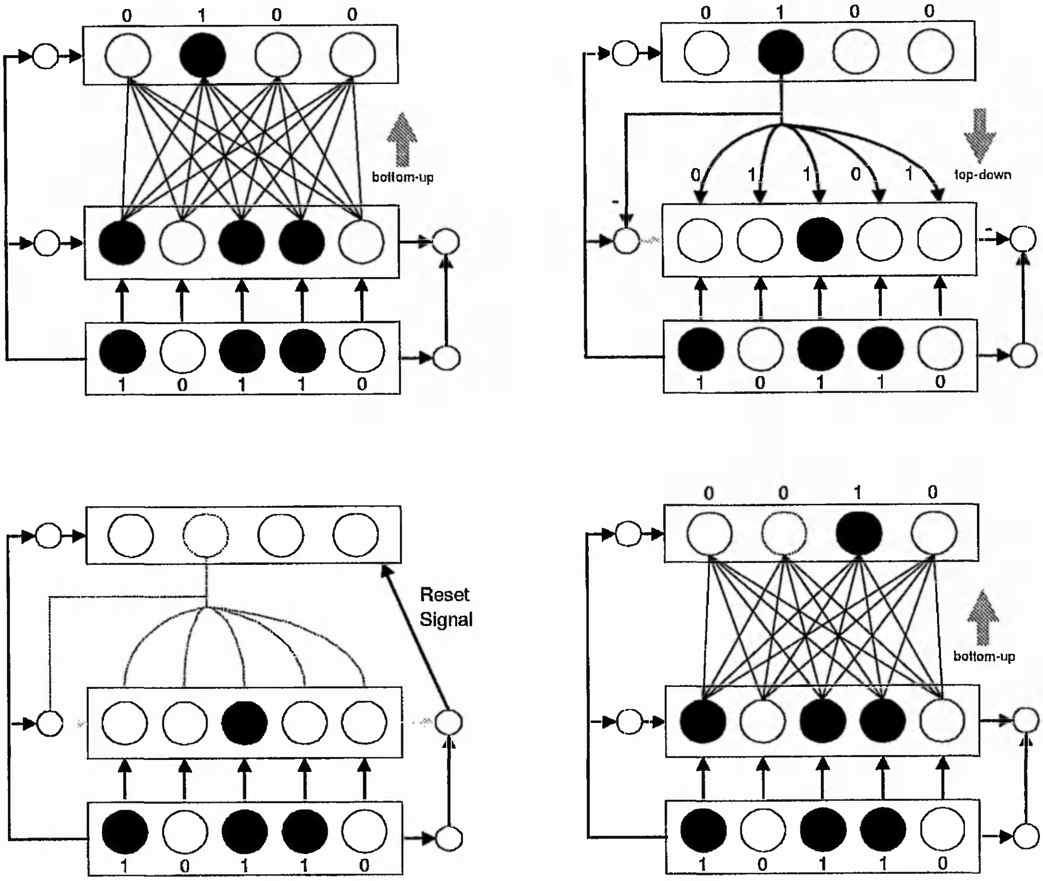

Figure 4.2: The ART search and hypothesis testing cycle: (a) an input pattern **I** = 1,0,1,1,0 registers itself as a pattern **X** across F_1. **X** is sent to F_2 via bottom-up weights and following competition an activity vector **Y** = 0,1,0,0 across F_2 emerges. (b) To test the hypothesis **Y**, the vector is sent back to F_1 via top-down weights resulting in a top down input vector **V** = 0,1,1,0,1 being fed to F_1. A matching function (in this case: $\mathbf{I} \wedge \mathbf{V} = \mathbf{X}^*$, as used in ART1 & Fuzzy ART) results in the a new vector $\mathbf{X}^*$ = 0,0,1,0,0 across F_1. (c) Because $\mathbf{X}^*$ is so small inhibition to the orienting subsystem from F_1 drops to a level whereby a reset signal is propagated to F_2. This inhibits the pattern **Y**. (d) **X** is again sent to F_2 via bottom-up weights leading to a new activity vector **Y** = 0,1,0,0 at F_2.

sufficiently matches **I**, (2) if **I** is too different from any of the previously learned category representations, an uncommitted F_2 unit is selected, which establishes a new category, or (3) the entire capacity of F_2 is utilized, in which case **I** cannot be accommodated.

Solving the Dilemma

Two particular mechanisms of ART processing that are instrumental in providing these architectures with their all important properties are:

- **Competition**: As a result of competition, knowledge within an ART architecture is localized[5], i.e. each F_2 unit becomes tuned to a single category (e.g. one's mother's face) and the sensory information associated with this category is stored across a *subset* of LTM weights, viz. those that connect all F_1 units to the particular F_2 unit in question. If the system is required to update information about this particular category (plasticity), then only the values of this subset of weights are adapted, all of the other weights in the system (i.e. those storing information about other categories) remain unchanged, hence, the stability. In other CMs that lack competitive mechanisms such as MLPs, the knowledge related to a particular category is not localized, it is fully distributed across all of the systems weights. Thus, updating information that relates to one category, alters all of the other categorical information in the system. New information overwrites old information.
- **Novelty Assessment**: Competition alone does not suffice to remedy the plasticity-stability dilemma. To explain further, imagine a CM that has three 'localist' category units. The first has already acquired information relating to the visual appearance of dogs, the second to the visual appearance of elephants, and the third has not yet undertaken learning, i.e. an uncommitted unit. The task at hand concerns acquiring information about wolves whilst not disrupting the other categorical information. As a result of the bottom-up activation and F_2 choice stage of processing, it is likely that

[5] It is worth noting that whilst localist connectionism has long been overlooked in favour of distributed connectionism, in recent years the former has been gaining some notable support, with a number of authors pressing to dispell some of the unfounded myths that surround localism (Thorpe 1995, Grainger & Jacobs 1998, Page 2000).

> the dog category will be chosen by competition as that which best matches the input – given the visual similarities between wolves and dogs. In a number of competitive CMs the processing cycle ceases at this point and thus the learnt dog information would be updated to take into account the new wolf information, thus corrupting the dog category[6]. However, in an ART system the choice of the dog category is only seen as a working hypothesis. The top-down matching process compares the hypothesised units categorical information to the input information and actually measures the extent to which they are similar. In this case, if the vigilance level is set appropriately, the wolf related input will be seen as too dissimilar, i.e. novel, *vis-à-vis* the dog related category information. The dog information remains unchanged, and following search, the third uncommitted category unit can learn the novel wolf information. Hence, plasticity and stability are achieved.

Vigilance is a vital element in ART processing in that it provides explicit control over the degree of generalization that a model exhibits. If vigilance is set low then many different types of inputs (e.g. dogs and wolves) are grouped together. Alternatively, if it is set very high ART simply becomes an exemplar learning machine; each input variation activates a different category unit.

As has been already noted, the ART processing cycle described above, is a feature shared by all ART CMs. However, each specific variation of ART also has its own distinct characteristics which stem from (1) its fan-in (or activation) function, (2) its matching function, and (3) its learning function. Two specific ART implementations are used later in this thesis; Fuzzy ART and Gaussian ART. The specific functions that each of these implementations employs are provided for reference in appendix A.

[6] The rules governing bottom-up activation and F_2 choice in ART systems are very similar to those found in two other classes of competitive CM: Kohonen's Self-Organising Maps (Kohonen 1982), and Rumelhart and Zipser's Interactive Activation models (Rumelhart & Zipser 1985) (as employed by Dorffner; see section 3.2.4). The lack of top-down processing and novelty detection in these other CMs has been directly linked to the instabilities in learning that each of the model types has been shown to exhibit (Grossberg 1976*a*, Grossberg 1987).

4.2.2 Supervised ARTMAP systems

The ARTMAP architecture was first introduced in 1991 by Gail Carpenter and colleagues (Carpenter et al. 1991) and is largely a fusion of two unsupervised ART modules[7]. The fusion of modules is accomplished by way of a third module – containing a field of units F_{ab} – which is able to learn a map that associates the categories learnt by one ART module with those learnt by the second ART module. Figure 4.3 illustrates the general form of a supervised ARTMAP model comprising two ART modules, ART$_a$ and ART$_b$; and a Map Field module . ARTMAP processing is described as supervised in the sense that if two patterns are presented to the model simultaneously then it is assumed that these patterns are linked. The model's task, given this direction, is to learn for itself the link between these patterns. The process is as follows[8]:

IV. Map Field Activation

Normal ART processing proceeds as described above up to a certain point, i.e. two resonant vectors $\mathbf{Y}_a$ and $\mathbf{Y}_b$ result in the choice of two category units J and K in the respective ART modules. At this point the normal cycle is broken in that long-term memory (LTM) learning does not occur. Instead the vector $\mathbf{Y}_a$ is sent via multiple fan-ins of adaptive weights resulting in an activity vector $\mathbf{Z}$ across F_{ab}.

V. Map Field Hypothesis Testing

An activity vector $\mathbf{Z}$ effectively establishes a hypothesis that an ART$_a$ category is connected to a certain ART$_b$ category. In order to test this hypothesis $\mathbf{Y}_b$ is transported to F_{ab} wherein $\mathbf{Y}$ and $\mathbf{Z}$ are compared[9]. If $\mathbf{Y}_b$ and $\mathbf{Z}$ are close enough, by some measure (a matching

[7] An ARTMAP model is just one class of ART model. An ARTMAP model also contains two ART models. The semantics of this can get confusing, thus in the context of an ARTMAP model an unsupervised ART model is referred to throughout th forthcoming text as an ART module.

[8] Further detail of ARTMAP processing can be found in appendix A.

[9] At this point, when receiving input from two sources, the dynamics of the Map Field are essentially the same as those operating at F_1 when it is receiving bottom-up and top-down input (see phase II of ART processing).

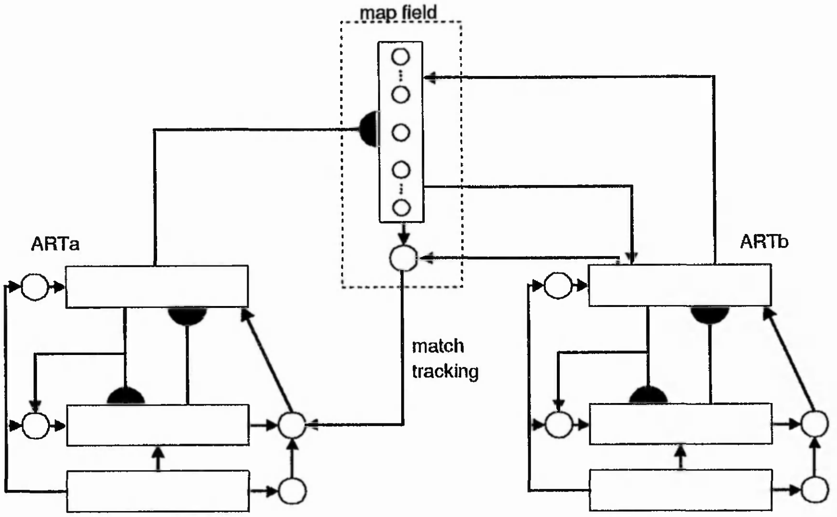

Figure 4.3: The diagram shows two unsupervised ART modules being linked together via a Map Field. This fusion results in the creation of a supervised ARTMAP model.

function, or rule), in relation to an ARTMAP system parameter ρ_{ab}, then the hypothesis is confirmed (a match) and resonance follows. If they are not close enough then the hypothesis is disconfirmed (a mismatch), and ARTMAP begins a search for a better 'hypothesis.'

VI. Resonance, or Match Tracking and ART$_a$ Search

Following a Map Field match learning ensues and the LTM weights in both ART modules and those connecting to F_{ab} may be adapted. Alternatively, following Map Field mismatch, a process called Match Tracking automatically sets ART$_a$ vigilance to a level whereby mismatch occurs in the ART$_a$ module. Normal ART module processing ensues (see phase III of ART processing) which leads to a new activity vector $\mathbf{Y}_a$, and the ARTMAP processing cycle starts again.

4.3 ART: An Alternative Framework for ENLP Models

One only has to do a brief scan on academic literature to find that ART and ARTMAP networks have found favour in a wide variety of practical applications throughout recent years. For example, in a Boeing parts design retrieval system (Caudell et al. 1994), robotic control (Dubrawski & Crowley 1994), and in several 3D object recognition systems (Seibert & Waxman 1992, Keat 1996). In this the last section of the chapter the case is put for the adoption of ART in a new practical context, viz. that of lexical acquisition and grounding (LAG) and the creation of applied ENLP models.

4.3.1 ART or ARTMAP

The first point to note is that whilst ART as a theory is suited to the LAG task, isolated ART models, of the unsupervised variety, are not. As discussed previously in section 3.1.3, a model that can only process a single type of input may be able to acquire or ground lexical symbols, however, it will not be able to acquire *and* ground such symbols. ARTMAP models on the other hand do provide a good fit with the generic LAG architecture described in chapter 3. This can be seen in the similarity between figures 4.3 and 3.3. Henceforth it is the ARTMAP architecture that is considered as a candidate for the LAG task.

4.3.2 ARTMAP and the 'Model' Model

The candidacy of ARTMAP is best assessed in relation to the properties one would expect an ideal model to exhibit. These properties, first introduced in section 3.2.1, shall now be considered in turn:

Learning Properties:

- *Autonomy* – The ARTMAP architecture has been specifically devised to process two independent sources of input data, categorise these data and then learn mappings between units that symbolically represent such categories. Once underway this process does not require intervention. Moreover, in contrast to many other CMs (e.g. MLPs) ARTMAP is initialized with ease. As far as topology goes one only has to consider the initial size of each ART module's F_0 layer, which is obvious given that one knows the type of patterns that are to be received by an embodied systems sensory transducers. ARTMAP does not require a designer to provide their own parameterisations for initial LTM weight values or to concern themselves with connectivity patterns between layers and processing units.

- *Flexibility* – An early ART implementation ART1 was confined to processing binary input patterns. However, this limitation has long been overcome and ARTMAP models based upon Fuzzy- and Gaussian ART are not constrained as such.

- *Adaptable & Scaleable* – As discussed in depth above ART was created from the ground up to solve the Stability-Plasticity Dilemma. The system is forever plastic and information can be added or adapted throughout a model's lifetime. ARTMAP also provides a proven scaleable architecture[10] which has been deployed in numerous data intensive applications (see, for example, (Carpenter et al. 1992, Williamson 1995)). ARTMAP was devised in order to process only two input channels. This provides limited scope for lexical development because lexical symbols gain meaning from multiple sensory groundings. This limitation is addressed in chapter 7.

- *Performance* – Competition in ART networks results in localised learning, i.e. only a very few weights require adaptation in order to code new knowledge. Moreover, ART networks in fast learn mode can stabily learn a new input pattern in relation to a sigle presentation of that input. These properties combined result in a system that is able to learn extremely quickly - typically orders of magnitude faster than non-localized CMs, e.g. MLPs.

[10] All of the ARTMAP systems described in the following chapters were implemented by the current author based on the theory discussed above and complying to the equations detailed in appendix A (see also adapted equations in chapter 7). The implementation was founded on object-oriented programming principles wherby each F_2 unit and each F_{ab} unit was created as an individual object. This allowed the author to create an adaptive architecture in which category units could be added as and when required and which always had a size optimal for the task at hand. Thus, in this case architecture scaleability is bounded only by memory limits on the host computer.

Prediction/Recall Properties

- *Description & Identification* – An 'out-of-the-box' ARTMAP system provides a framework to enable either identification or description dependant on which ART module processes which type of input. This is because the fan-in of LTM weights that connect to the F_{ab} layer were originally designed to only provide a one-way transport for layer activity. However, only a simple algorithmic modification is required to enable two-way transport via these LTM weights and thus enable both identification and description to be undertaken.

- *Performance* – Recall performance is not really an issue in an ART-based architecture as Grossberg notes: "It uses direct access to recognize familiar events [patterns] with a speed as fast as one's hardware can run." Even as an architecture scales[11] input patterns can rapidly be categorised.

In summary, one hopes that this chapter has provided the reader with an adequate rationale as to why ART-based networks might provide computational framework better suited to the task of lexical acquisition and grounding. The three chapters that follow seek to substantiate this rationale through practical investigation.

[11] In a software based serial implementation there is a computational cost involved with increases in F_2 size as each F_2 unit must integrate the input it receives (fan-in function) in turn. Thus a size increase will result in a performance hit *vis-à-vis* overall processing times. This would not be the case for a parallel implementation.

Chapter 5

Investigations with ART, I: Colours and Categories

In the last chapter Adaptive Resonance Theory was introduced and shown to offer a number of beneficial qualities over many other connectionist processing methodologies. It was thus proposed as a superior candidate model with which to acquire and ground lexical symbols. This chapter considers how one might move towards creating a model that could understand and interact with humans as they talk of colours, i.e. a model that could acquire and ground colour terms.

The aims of the work presented in this chapter are twofold:

1. To begin to demonstrate to the reader the validity of the claims made *vis-à-vis* ART-based models and the LAG task.

2. To demonstrate the importance of any LAG model's 'window-on-the-world'. By virtue of the nature of the transduction equipment at a model's sensory peripheries it is presented with a certain view of things. Harnad has argued that such representations should be *iconic* (Harnad et al. 1994). This argument is discussed and considered further in relation to the new Triangle of Meaning (TOM) which was introduced in the latter sections of chapter 2.

5.1 The Colour Task

To successfully acquire and ground colour terms one would of course require a model somewhat akin to that specified in chapter 3 – a model able to display its understanding by way two recall abilities:

1. *Description* – Shown an example of a colour, the model should be able to name that colour by recalling the correct colour term, e.g. the symbol 'red'.
2. *Identification* – Provided a colour term, the model should be able to recall a colour that the term adequately describes, e.g. point out an example of red.

Whilst the ARTMAP models described later in the chapter are quite capable of performing such tasks (as shall be shown in the next chapter) in this chapter such a complete LAG task is not considered. Instead, the investigations presented in this chapter concentrate on how the representation of experience perceived at the sensory level of description can aid or hinder the task of learning the categorical concepts which serve to ground internal symbols in a LAG model.

Stevan Harnad first introduced the notion of *iconic* representations in his seminal 1990 paper on Symbol Grounding (Harnad 1990). Such representations he described as being "internal analog transforms of the projections of distal objects on our sensory surfaces" (Harnad 1990, p.342). In a later paper the importance of 'iconicity' was stressed and its sense further qualified as pertaining to "how analog, non-arbitrary, or structure-preserving [a representation] was in relation to what it represented" (Harnad et al. 1994). In light of the new TOM introduced in section 2.5.1, Harnad's qualification can become a little confusing. One may remember that in the new TOM, objects do not exist out there in the world. This begs the question: In relation to what exactly should an iconic representation be structure-preserving?

The answer offered here is that 'iconicity' in the context of the new TOM means that representations employed at the sensory level of one body should preserve some structure of those employed by a second body (see figure 5.1). Furthermore, it is suggested that iconicity is important in this sense in that it provides a similar substrate for categorisation

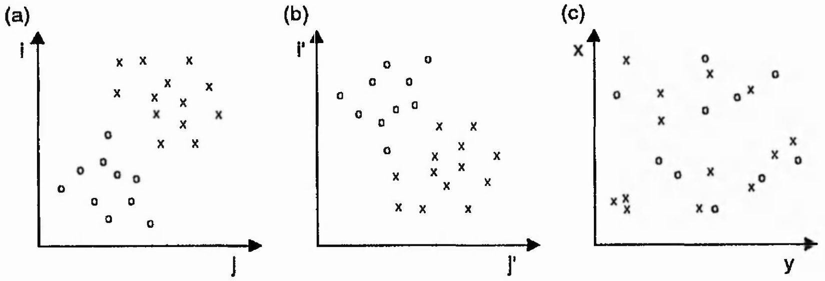

Figure 5.1: (a) shows details how one body transduces some aspects of the outside world in terms of components i and j. (b) shows how a second body transduces the outside world in terms of components i' and j'. (c) shows how a third body transduces the outside world in terms of x and y. The second body preserves to a large degree that which is represented by the first body, and vice versa. The third body does not preserve any such structure.

and hence the two bodies should be able to develop a similar categorical level understanding of things with greater ease.

Hence in building artificial models that are to understand the world in a similar way to humans – in order to communicate about this world – it is suggested here that one needs to take advantage of what understanding there is of how humans process and organise their input stimulus due to their physical embodiment. In doing so it is proposed that one provides cues that will improve the learning function, i.e. building categorical concepts upon such representational substrates will be easier than doing so upon representations that are arbitrarily crafted by a designer. It is this theory that is put to the test in the investigations that follow.

ARTMAP based models paired with the domain of colour provide a means by which to test this theory. Firstly, the domain of colour has been chosen because colour science is an established research area where intensive investigations have considered the cognitive aspects of colour processing in humans. This work has spawned a body of theories that consider how colours may be cognitively represented and that have produced various *colour spaces*, *colour models*, and *colour order systems*. These representational models are based upon both physiological and psychological data and thus adhere to a set of human specific constraints. Secondly, ARTMAP based models provide a number of means by which one can assess how easily a specific conceptual quantization of the input feature space – to provide colour categories – can be obtained. Faster learning rates, better levels of performance

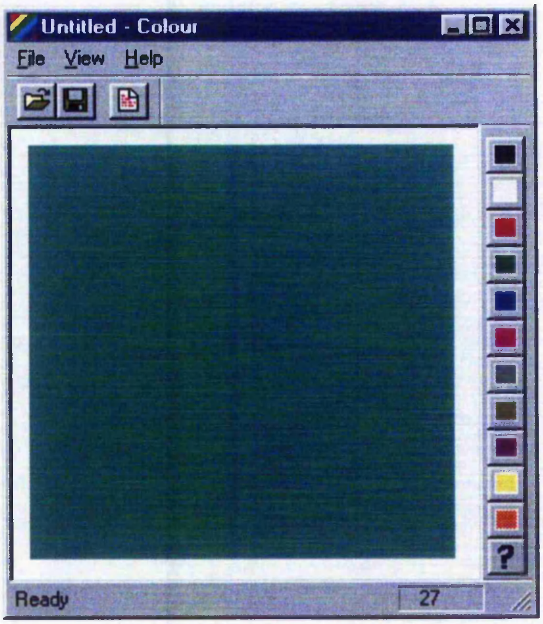

Figure 5.2: Data Collection Application

when presented with novel data, and smaller populations of F_1 level units are all measures which suggest a particular ARTMAP model required less computational power to achieve a certain level of performance.

5.2 Data Collection

The first step of the investigations required that named colour data be collected to provide input data for the ARTMAP learning system. A simple colour naming application (shown in fig 5.2) was developed to enable this data collection. The application generated a random solid block of colour and displayed it on screen. A human subject was then simply asked to categorise each block of colour as it appeared by pressing one of eleven buttons, each of which related to one of the following eleven colour terms:

{*black, white, red, green, yellow, blue, brown, pink, purple, orange, grey*}

These eleven terms have been defined by Berlin & Kay (1969) as being the *basic colour terms* used by English speakers. Basic colour terms have a common set of characteristics (Lammens 1994) in that :

1. They are "monolexemic"; a term's meaning is not predictable from the meaning of its parts e.g. green and yellow are monolexemic, whereas greenish is not.
2. Their "signification" does not fall under that of any other colour term, e.g. crimson, scarlet both fall under the term red.
3. They apply to a wide class of objects, e.g. blond would not qualify as it generally only applies to the description of hair.

The generation of a colour sample was achieved by the random choice of a red (R), green (G), and blue (B) component from within an *RGB* colour cube, where each component had a linear range varying between 0 and 255. Therefore, each sample was chosen from a possible set of 16,777,216 samples. However, for display purposes, a 16 bit palette was used thus reducing the set of possible samples to the order of 2^{16}, or 65,536 colours. The *RGB* model was used for the initial data collection for two reasons; firstly it is the main colour model supported at the software application programming level, and secondly it is based upon physiological constraints found within the human retina and can thus be considered as a prime candidate for a non-arbitrary structure-preserving representational model.

For the purpose of input to the ARTMAP system, each colour term chosen was represented by one of the eleven orthogonal vectors shown in table 5.1.

The initial data collection process produced a named set of 1200 *RGB* colour samples. From these data three separate train/test sets were produced, each training set consisting of 1000 colour/linguistic vector pairs drawn randomly from the named sample set with each test set comprising the remaining 200 vector pairs. Each *RGB* colour component was scaled from a range 0f {0 - 255} down to a range {0 - 1}[1] thus a typical vector pair taken from an *RGB* based training set was as follows :

0.607843 0.419608 0.364706 0 0 0 1 0 0 0 0 0 0 0

The next stage required that the named *RGB* based data sets be transformed into data sets based on alternative representation schemes. To test the theory that *iconicity* is important, meant creating some transforms that did little to preserve any of the spatial structure found in the *RGB* model and some that did. Herein, the former are described as being arbitrary

[1] the allowable adaptive resonance input range

orange	1,0,0,0,0,0,0,0,0,0,0
black	0,1,0,0,0,0,0,0,0,0,0
blue	0,0,1,0,0,0,0,0,0,0,0
brown	0,0,0,1,0,0,0,0,0,0,0
green	0,0,0,0,1,0,0,0,0,0,0
grey	0,0,0,0,0,1,0,0,0,0,0
pink	0,0,0,0,0,0,1,0,0,0,0
purple	0,0,0,0,0,0,0,1,0,0,0
red	0,0,0,0,0,0,0,0,1,0,0
yellow	0,0,0,0,0,0,0,0,0,1,0
white	0,0,0,0,0,0,0,0,0,0,1

Table 5.1: Vector Coding of Eleven Basic Colour Terms

representations, the latter as non-arbitrary.

5.2.1 Converting *RGB* to *HSV*

The Hue (H), Saturation (S), Value (V) colour model is part of a larger family of colour models that are based on psychological and psychophysical findings centred around the appearance of colours to human observers. *HSV* was therefore classified as another example of a non-arbitrary representational model. The model is based upon a hexacone co-ordinate system with the hue, saturation, and value dimensions being roughly analogous to the dimensions on tint, shade, and tone as used by artists. In order to obtain an *HSV* based data set, a conversion process was employed; a point in the *RGB* colour cube was transformed to a point in the *HSV* hexacone via the sequential application of equations 5.1 through 5.8 shown below.

$$V = max(r, g, b) \tag{5.1}$$

$$temp = min(r, g, b) \tag{5.2}$$

$$S = \begin{cases} 0 & \text{if } V \equiv 0 \\ \frac{V-temp}{V} & \text{otherwise} \end{cases} \tag{5.3}$$

$$temp^r = \frac{V-r}{V-temp} \tag{5.4}$$

$$temp^g = \frac{V-g}{V-temp} \tag{5.5}$$

$$temp^b = \frac{V-b}{V-temp} \tag{5.6}$$

$$H = \begin{cases} undefined & \text{if } S \equiv 0 \\ temp^b - temp^g & \text{if } V \equiv r \\ 2 + temp^r - temp^b & \text{if } V \equiv g \\ 4 + temp^g - temp^r & \text{if } V \equiv b \end{cases} \tag{5.7}$$

$$H = \begin{cases} (H \times 60) + 360 & \text{if } H < 0 \\ (H \times 60) & \text{otherwise} \end{cases} \tag{5.8}$$

The conversion was applied to each of the three RGB train/test sets. In each case the hue component produced by the conversion was rescaled from a {0 - 360} range down to a {0 - 1} range. The encoding of lexical items was kept as per the RGB data file i.e. eleven bit binary vectors (table 5.1).

5.2.2 Converting RGB and HSV to RGB^* and HSV^*

RGB^* and HSV^* were devised as representational models of colour which could maintain the full range of input allowable within both the original RGB and HSV models and yet allow all data to be encoded in a compact binary form.

The RGB^* data was obtained by transforming the three numeric dimensions of the original RGB data into twenty four binary valued dimensions. Each numeric RGB dimension was re-scaled to the {0-255} range and then mapped to an eight bit binary vector. A typical *green* RGB vector $\{16, 191, 86\}$ was transformed as follows :

The HSV^* data was obtained via the same principle but was derived from the HSV numeric data, rather than the RGB data. Again the encoding of lexical items was kept constant in

$$\begin{array}{lcl} 16 & \longrightarrow & 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0, \\ 191 & \longrightarrow & 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 0,\ 1, \\ 86 & \longrightarrow & 0,\ 1,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0 \end{array}$$

both cases (table 5.1).

The conversion process used in each case above was highly non-linear in nature. It warps the representational space, leaving some points that were close to each other in the *RGB* model as distant relatives. Because of this, RGB^* and HSV^* were considered as examples of arbitrary representational models.

5.2.3 Converting *RGB* to RGB^{**}

RGB^{**} was the final representation scheme used to encode the data. As with the models described in the preceding section, a binary based encoding of data was employed that allowed the full range of the data from the original *RGB* named data set to be maintained.

To obtain the RGB^{**} data, each dimension of the original *RGB* data (again re-scaled to {0 - 255}) was mapped to a 256 bit binary vector; the vector elements having a value of 1 if their position within the vector was less or equal to the numeric value being transformed, and having a value of 0 if all other cases. Thus, the red element of the *green* *RGB* vector $\{16, 191, 86\}$ would be mapped to a 256 bit vector having the first 16 bits set to 1 and the remaining 240 bits set to 0. The overall encoding vector for each colour within RGB^{**} data set was thus quite large, i.e. a 768 bit binary vector.

The conversion from *RGB* to RGB^{**} has a linear nature, therefore the structure of the data was kept intact and RGB^{**} was thought of as another example of a non-arbitrary representational model.

5.3 Creating Colour Categories

The data collection process provided a total of 15 train/test sets, 3 each for the 5 representation schemes introduced in the previous section. In order to investigate the effect each representational model made to the grounding task, two ARTMAP variations were used: Fuzzy ARTMAP with complement coding (Carpenter et al. 1992), and Gaussian ARTMAP (Williamson 1995). By using the two network variations it was possible to additionally assess how the different competitive learning metrics used in each system (see appendix A) effected the overall task.

ART_a Base Vigilance, $\overline{\rho_a}$	=	1.0 x 10^{-307}
ART_b Vigilance, ρ_b	=	1.0
Alpha, α (Fuzzy)	=	0.1
Gamma, γ (Gaussian)	=	1.0

Table 5.2: Parameters used in Grounding Investigations

For each data set, the training portion of the set was applied to each ARTMAP variation for a total of twenty epochs, after each epoch two measures of performance were applied; the population of ART_a categories was determined and the error on the previously unseen test portion of the data set calculated. A final measure of performance, the total processing time taken to complete the 20 epoch train/test cycle was recorded for each data set. Table 5.2 provides an account of the settings of the ARTMAP free parameters used in all of the trials. All of the investigations were carried out on a 100MHz Pentium PC with 16Mb main memory.

5.4 Results and Discussion

Accounting for variations in the ARTMAP model and colour representation used, a total of 30 result sets were obtained from the grounding investigations. In this section these results are introduced and discussed in detail.

Three data sets were produced for each specific colour representation used in order that the results obtained could be averaged to provide the underlying statistical consistency.

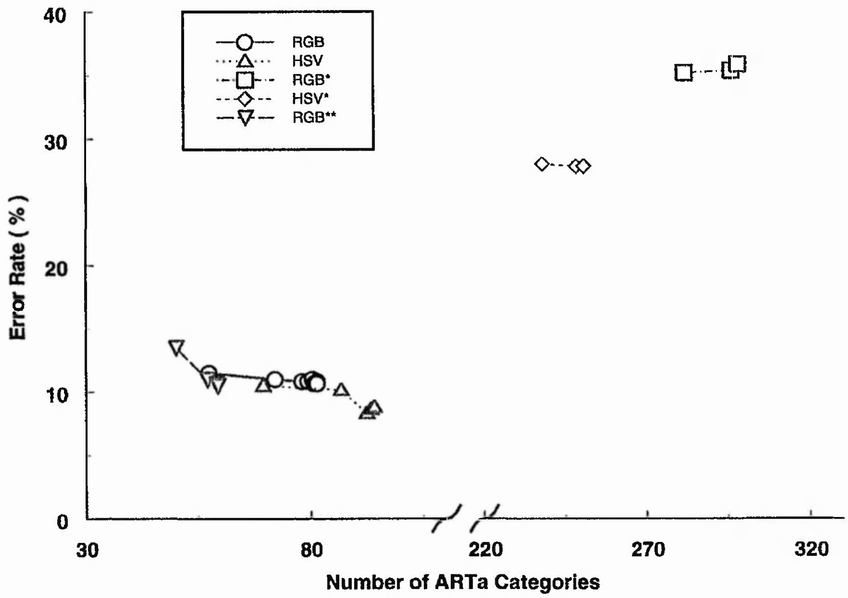

Figure 5.3: Fuzzy ARTMAP with Complement Coding.

Therefore, for each ARTMAP variation five averaged result sets were obtained, one for each of the representational schemes used.

Figure 5.3 shows the averaged results for grounding in the Fuzzy ARTMAP model. The results strikingly show that the RGB^* and HSV^* based models produced a significantly greater population of ARTMAP categories; $\approx$ 200% more than the other models. Also evident from the graph are two distinctive groupings of activity, the RGB^* and HSV^* trial results are placed to the top right (TR) whilst the RGB, HSV, and RGB^{**} results appear to the bottom left (BL).

A similar pattern of results, shown in figure 5.4, were also obtained from the Gaussian ARTMAP trials. Again, the arbitrary RGB^* and HSV^* based models exhibit a category count that far exceeds those seen in the other models and once more there is some evidence of a BL / TR separation of data.

The BL / TR separation supports the idea that a bad representational input model will cause an ARTMAP system to proliferate categories. A large comparative amount of categories conveys that the learning system has an inferior overall idea of the general mapping that

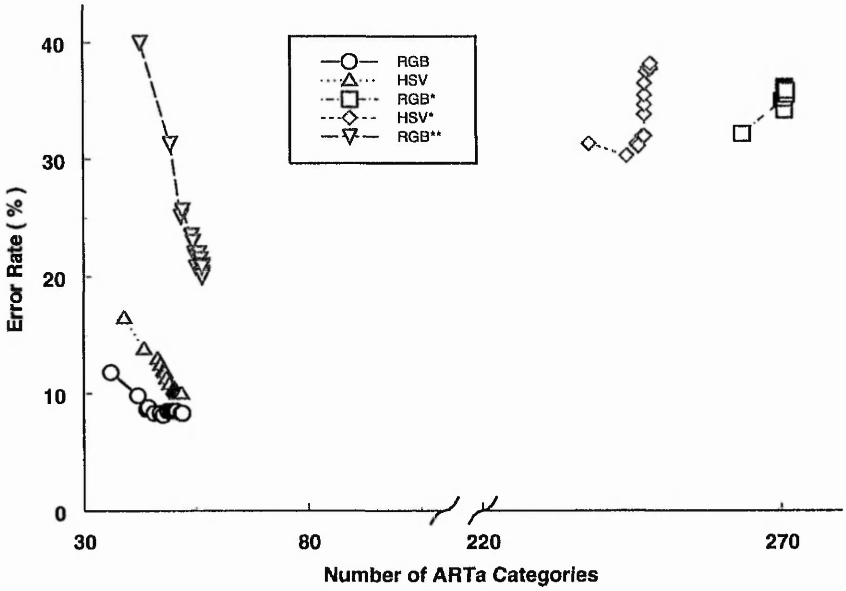

Figure 5.4: Gaussian ARTMAP.

exists between the input and output data, i.e. a poor generalisation. Therefore, one would expect, as is seen, that when tested on novel or unseen data, the predictive abilities of a system with a large category count would fall short of a system with fewer categories.

Each graph quite clearly demonstrates a performance difference between those colour models that were preconceived as having a non-arbitrary nature, and those that were arbitrarily devised. For each type of learning methodology used, the embodied representations of colour require less supervised re-categorisation and allow the grounded systems in which they are used to correctly predict the output classes for novel test data to a far greater degree.

In each of the experimental trials undertaken, the processing times required to train and test each ARTMAP system for a 20 epoch cycle were recorded. Figures 5.5 and 5.6 provide these results in a graphical form. The most striking result that can be seen for both the Fuzzy and Gaussian based systems, is the considerable computational cost on both learning and recall when the RGB^{**} is used. This cost is quite understandable due to the fact that the number of weight based calculations increase proportionally with the dimensionality of the input vectors used in all connectionist systems. This proportionality is much higher in

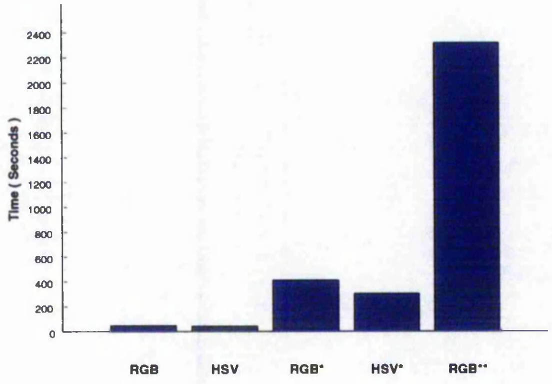

Figure 5.5: Fuzzy ARTMAP processing time. The figure shows the processing times required by the Fuzzy ARTMAP model to complete a full twenty epoch train/test cycle for varying colour representation models.

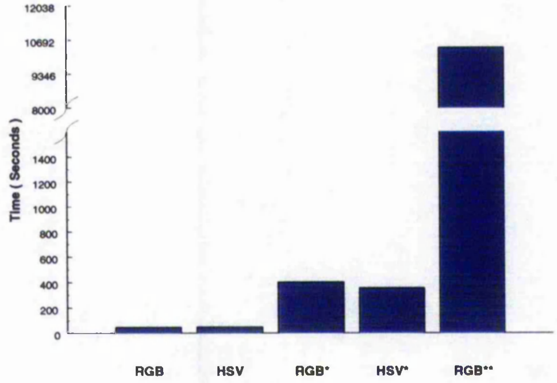

Figure 5.6: Gaussian ARTMAP processing time. The figure shows the processing times required by the Gaussian ARTMAP model to complete a full twenty epoch train/test cycle for varying colour representation models.

models that employ fully distributed learning such as the MLP network family. Therefore, one would expect the RGB^{**} data encoding method to slow down grounding systems such as those described in sections 3.2.3- 3.2.5, and 3.2.8 to a far greater degree than is shown in the ARTMAP results.

Although both ARTMAP models suffered a delay in processing the RGB^{**} data, the Gaussian based model can be seen to be significantly slower than its Fuzzy counterpart. In the worst case a Gaussian based system processed the RGB^{**} data 24 x 10^3 percent slower than when using the RGB data. The Gaussian based model also obtained considerably poorer % error test results, when trained on the RGB^{**} data than did the Fuzzy model.

This disparity can be explained by an inconsistency between the very high dimensional input used in the RGB^{**} representation and the nature of the hyper-dimensional ellipses used to fit the data to categories in the Gaussian model. The binary nature of the data means that data points are clustered at the maximum and minimum of each dimension. For this reason the data is more suited to the multidimensional rectangles employed in the Fuzzy model. Williamson has clarified this point remarking that the hyper-rectangle "is perhaps best suited to data that are uniformly distributed within hyper-rectangles" (Williamson 1995, p.4).

An appraisal of the results presented so far, suggests, that for reasons of performance and computational cost, the RGB and HSV are best placed to practically implement colour grounding. Figure 5.7 compares how each of these models performed in each ARTMAP system. Overall, these results show that the Gaussian / RGB model performs best with the lowest error rate measured on test data, $\approx$ 8%, for a low category count, $\approx$ 50 categories. The number of categories directly effects the memory requirements of a software ARTMAP implementation, therefore, as a secondary consideration, it is computationally efficient to give implementational preference to those models which produce low category counts.

5.4.1 An assessment of failures

The best result obtained from the experimentation detailed in the previous section, was accomplished by grounding the second RGB data set using a Gaussian based model. Using this combination, an error rate of 6.5% was achieved after twenty epochs of training. Thus,

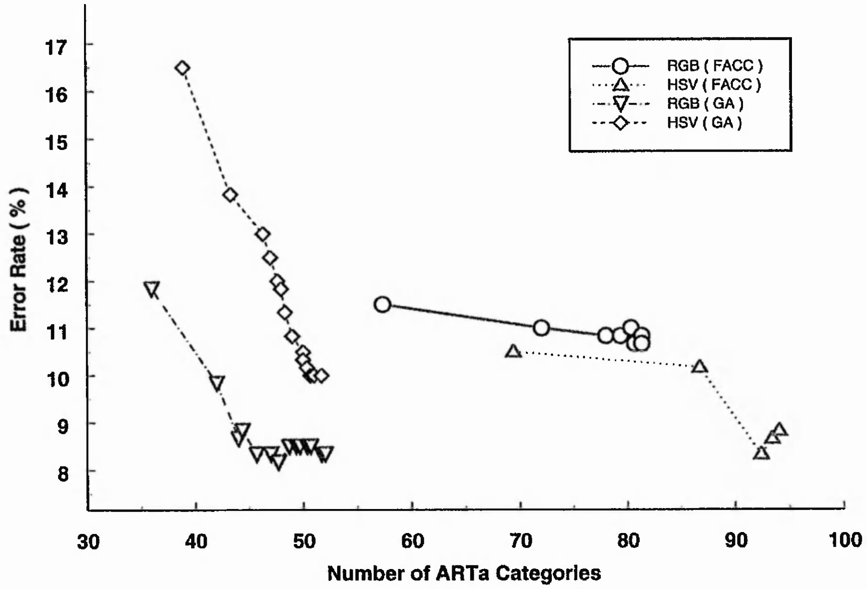

Figure 5.7: Fuzzy ARTMAP vs. Gaussian ARTMAP.

prediction failed for 13 patterns in the test set. To obtain a further insight into the nature of the failures, the trained system that produced each of the failure patterns was scrutinised.

Table 5.3 provides details of the failures in the second RGB test set. It shows each patterns R, G and B values, and indicates the colour name predicted by the model set against the name provided in the test set. Upon closer inspection a total of 8 of the 13 failures were found to be *borderline colours* (see figure 5.8), i.e. when reconsidered by a human judge they were found to be best described by a colour mixture, e.g. pattern 15 was a purplish blue. Further investigation highlighted that the colour predicted by ARTMAP was the complementary colour of the mixture in each of these eight failures. Only three failures were found to be significantly unsatisfactory.

5.5 Chapter Conclusions

This chapter has introduced two variations of connectionist models, both based upon Adaptive Resonance Theory. These models have been applied to a subset of a full LAG task, the

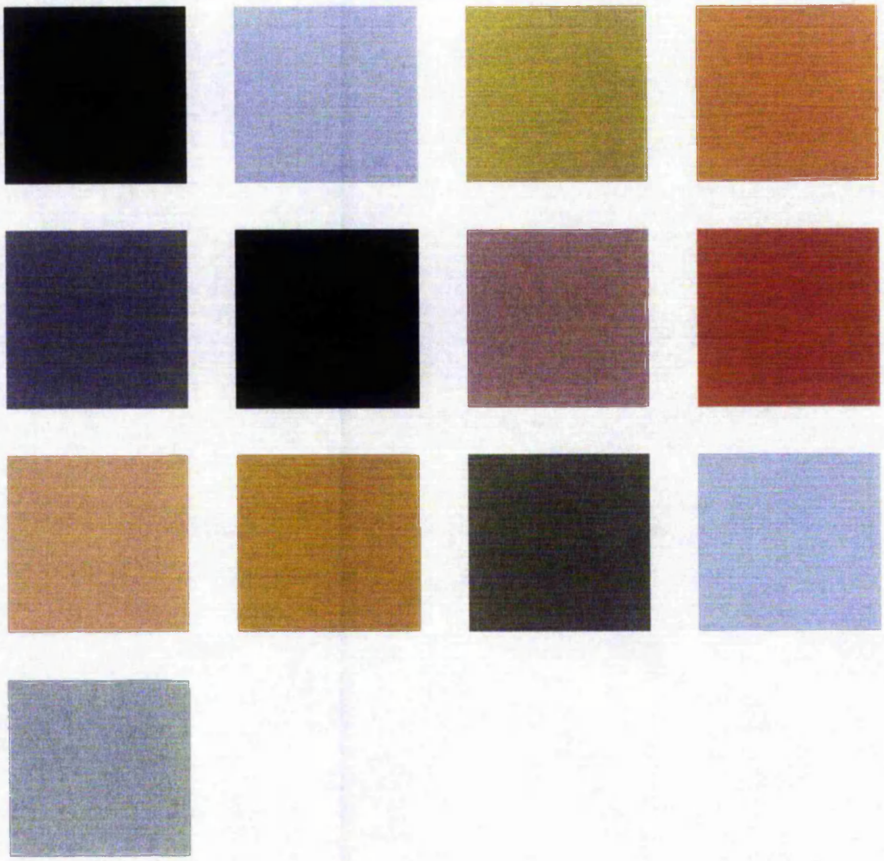

Figure 5.8: Failures. (1^{st} row - failures 1-4, 2^{nd} 5-8, 3^{rd} 9-12, 4^{th} - 13). Note: Blocks 1 and 6 have been shown as black blocks since the actual colours of these block are beyond the gamut of a printer.

Test Set Pattern No.	**RGB Value**	**ARTMAP Prediction**	**Name in Test Set**
15	134, 70, 250	purple	blue
47	188, 197, 217	blue	grey
53	196, 181, 19	green	brown
54	217, 147, 48	orange	brown
60	119, 112, 154	grey	purple
80	206, 10, 126	red	pink
86	184, 135, 115	grey	brown
98	204, 91, 52	brown	orange
119	217, 174, 100	green	brown
139	208, 165, 25	orange	brown
141	124, 137, 57	brown	green
178	194, 204, 223	blue	grey
184	169, 187, 165	green	grey

Table 5.3: The table above provides details of *RGB* based colour vectors that were incorrectly described by the ARTMAP models (see main text for further details).

task being to learn to describe colours by way of a number of symbolic tokens. The results have in a number of cases provided evidence that ARTMAP models are able to succeed in learning this task. One Gaussian ARTMAP model was found to provide the greatest performance. Accounting for borderline colours the system was found to attain a fair/good prediction rate of 98.5%. In this respect the qualities of ART models alluded to in the previous chapter have not been challenged and one of the aims of the chapter has been met.

With respect to the second aim. It has been shown clearly in the empirical investigations above that representation at the sensory level is an important factor. Moreover, the results presented suggest that Harnad's call for iconicity was well placed. In all cases those models that formed colour concepts on the basis of structure-preserving, non-arbitrary representations were able to do so with far greater ease and success than those models whose view of the world was based upon arbitrary designer crafted representations.

Chapter 6

Investigations with Adaptive Resonance, II: A Comparative Study

The last chapter introduced a partial model of lexical acquisition and grounding in order to investigate the issue of representation. Since the colour related datasets employed were entirely novel, these simulations provided little scope to investigate how ARTMAP systems compare in an embodied semantic capacity to any of the related models introduced and discussed in chapter 3.

In this chapter a number of empirical investigations are undertaken that do provide such scope for comparison. These investigations are directly related to those undertaken by Sales (1996) (see sec. 3.2.8) in his work with WANN-based lexical acquisition and grounding systems, and also bear some resemblance to the investigations undertaken by both Plunkett et al. (1992), and Dorffner (1992) (see sections 3.2.5 and 3.2.4 respectively). Results are provided that further substantiate the claim that the new ART-based lexical acquisition and grounding proposal put forward in this thesis is indeed superior to the alternative contemporary proposals reviewed in chapter 3.

6.1 Acquiring Grounded 'Concrete' Nouns

The work that is described in this chapter centers upon the acquisition of grounded *concrete* nouns[1] which is a task that has been studied previously by Sales (1996). This particular task has been selected for two reasons:

1. As has already been discussed in section 3.2.8, the work reported by Sales has close parallels to the new work that is the subject of this thesis. Whilst theoretically the work is founded upon a body of 'human' sciences data (e.g. from Linguistics, Cognitive Science, and Psychology) that addresses the question of how humans have developed their languages and how these languages work, the primary aim is not to add to this particular knowledge but rather to *practically* apply what has already been learned in order to enhance human-machine interaction, e.g. to create robotic systems that can act upon and learn more about their operating domains through English dialogue.

2. Of all the work reported in chapter 3, the work of Sales is the only piece that provides a strong methodological description and analysis of the investigations that were undertaken, the data that were employed, and the results that were achieved. In all other cases such descriptions were either vague or under-specified, and as such provide no grounds from which to undertake any realistic empirical comparative analysis.

6.1.1 The Training Data

To re-cap briefly from section 3.2.8 the task of acquiring grounded concrete nouns as specified by Sales centers upon two streams of input: visual and linguistic. A total of 37 nouns are employed in the investigations, each being coupled to a pictographic image that provides a rough visual portrayal of the shape of the entity to which the noun refers. The 37 images employed are shown in figures 6.1 and 6.2 in the 144 bit (12 x 12) form that was used as visual input where each black square represents a value of 1 within the input vector, and each white square represents a value of 0.

[1] In this context the term 'concrete' is employed to describe a class of nouns that primarily refer to tangible physical entities, for example, nouns such as 'dog', 'chair', and 'cup'. Alternatively, nouns such as 'happy', or 'love' have a more abstract reference and cannot be described as concrete.

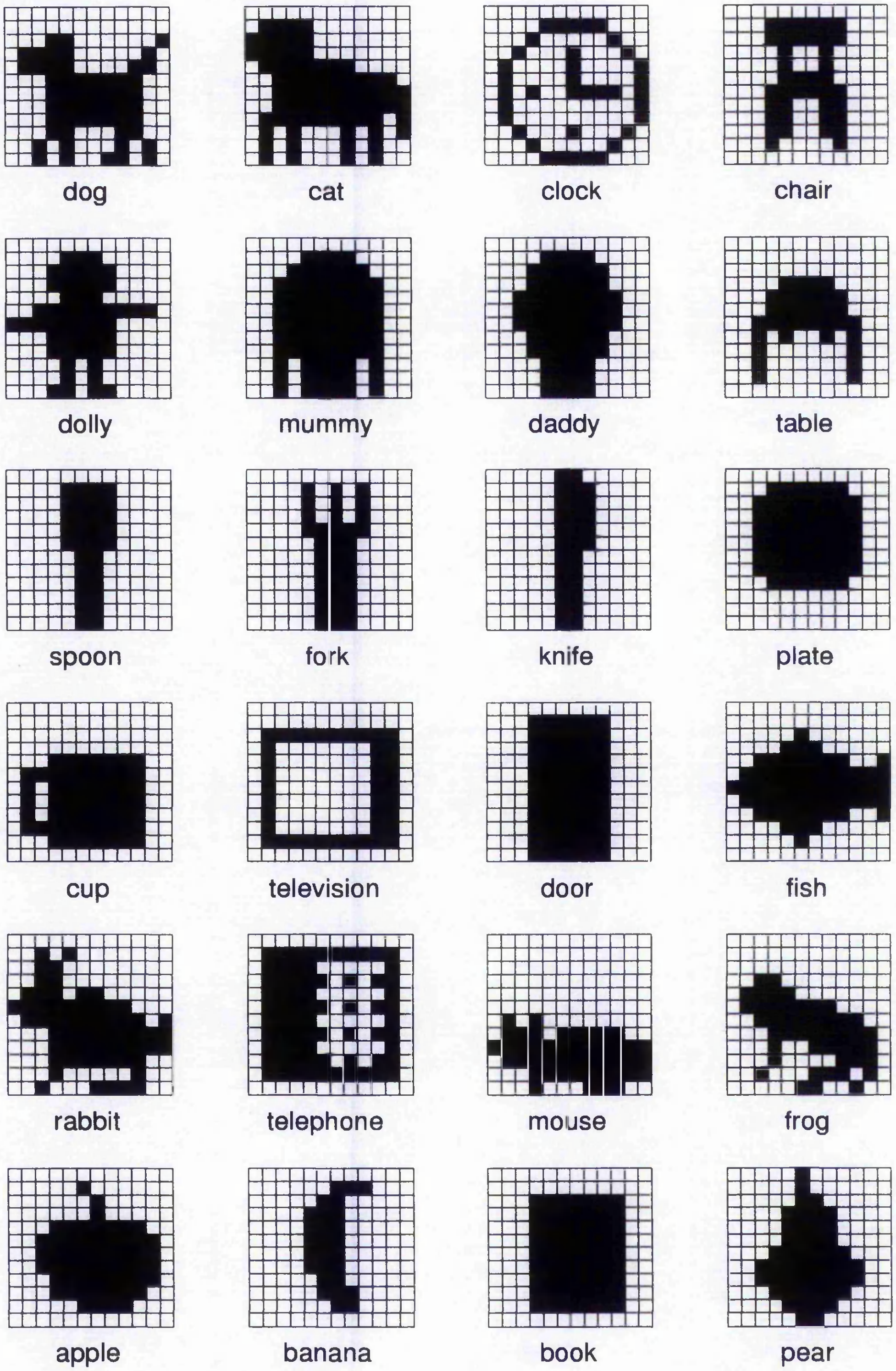

Figure 6.1: Patterns 1-24 (ordered left to right, top to bottom) of the 37 144-bit visual training patterns used by Sales in his lexical acquisition and grounding simulations.

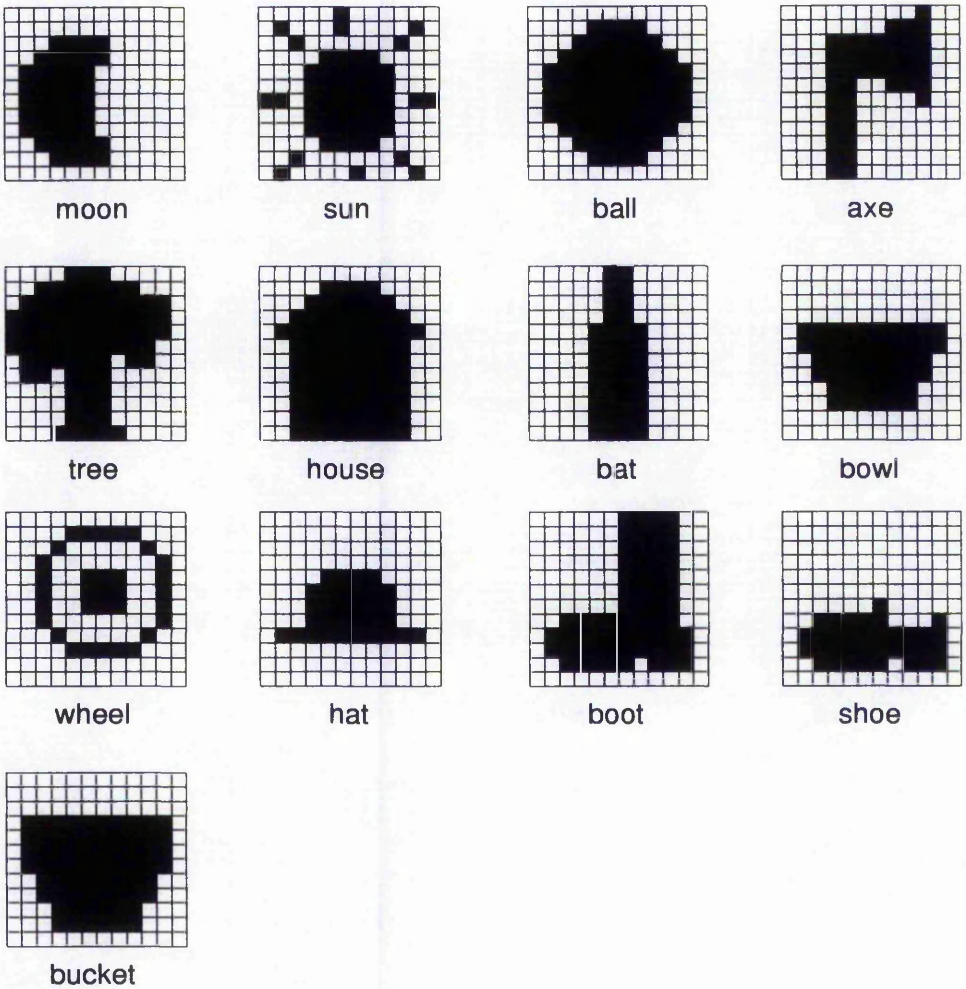

Figure 6.2: Patterns 25-37 of the 37 144-bit visual training patterns used by Sales in his lexical acquisition and grounding simulations.

The linguistic input is also encoded as a binary input vector. Each noun is presented as a 64 bit vector using an alphabetic encoding system which represents each letter within a word as a 5 bit binary value according to its position in the English alphabet, i.e. the letter 'A' has an associated value of 1, through to the letter 'Z' which has an associated value of 26. Hence, the word 'dog' is alphabetically encoded by the following 64 bit vector:

$$dog = \{\ 0010001111001111111111111111111111111111111111111111111111111111\ \}$$

The five leftmost bits (00100) represent a value of 4 in binary thus coding the letter 'D', the next five (01111) represent 15 (O), and the next five (00111) represent 7 (G). The remaining bits in the vector are all set to a value of 1 and act as padding. Of course words that contain a greater number of letters employ these other bits. Employing 5 bits per letter, the 64 bit encoding scheme is able to encode any word that has 12 or fewer letters. The last four bits are always padding and thus always have a value of 1.

6.1.2 Learning the Training Data

The nature of the training data engineered by Sales and described above is quite different to that described in the previous chapter. In the previous chapter the data employed consisted of many differing input vectors, or examples of each colour category to be learnt and the learning process effectively picked out a 'prototype' for each category. By contrast the training data introduced above only provides a single input vector for each object/entity category. Each such vector is in effect an *a priori* 'prototype' — the perfect categorical example.

In keeping with Sales original investigations and using such training data one can question whether the use of the term *learning* (in the sense described in section 3.1.1) is appropriate to describe the task that is being undertaken. In the context of an ARTMAP system, given that there is only one 'perfect' example of each pattern within the training data, there is no need for the system to utilize its generalization abilities. By initializing the vigilance values in the two ART modules to their respective maximum values,

ART_a Base Vigilance, $\overline{\rho_a}$	=	1.0
ART_b Vigilance, ρ_b	=	1.0

the ARTMAP system effectively becomes a template coding mechanism where rather than bringing the weights associated with each F2 level unit into equilibrium based upon a cluster of patterns, each pattern is simply coded by a unique F2 level unit, i.e. the weight vector associated with an F2 unit is a direct copy of the pattern vector it has coded. Hence, given the above data one would expect an ARTMAP system to develop 74 F2 level units in total. Half of these in the F_2^a layer, responsible for representing the 37 visual input patterns, and half in the F_2^b layer, responsible for the 37 lexical patterns. Moreover, despite the fact that the input vector dimensions are generally larger than those employed in the colour related investigations, the fact that ARTMAP can operate in this template mode, coupled to the fact that the size of the data set is notably reduced (1200 training examples were used in colour tasks, here only 37 are employed), one would expect to observe relatively quick training times.

All in all, three groups each containing five identical ARTMAP models[2] were trained using Sales' data. The visual input was processed by the ART$_a$ module and the lexical input by the ART$_b$ module. The first two groups were comprised of fuzzy ARTMAP models. The first group of models used the standard FA algorithm, whilst the second employed the additional complement coding normalisation technique (see appendix A, section A.1.3). The third group comprised 5 Gaussian ARTMAP models. In each case the models were trained through one epoch of the data (37 picture-word pattern pairs). The pairs were presented in a sequential order as shown in figures 6.1 and 6.2. To provide a basic initial analysis of the training process, each model was tested using the same training tests:

1. Given a visual input pattern in isolation each system was assessed to determine whether recall of the correct visual category was achieved.

2. Given a visual input pattern in isolation each system was assessed to determine whether recall of the correct lexical category was achieved (description).

3. Given a lexical input pattern in isolation each system was assessed to determine

[2] Five of each model were required for the testing phase (see below) where five different data sets were utilized. As with the colour related tasks discussed in the previous chapter all of the ARTMAP models were implemented by way of unoptimised executable C++ code, and all of the empirical investigations were carried out using a 100MHz Pentium computer with 16Mb of main memory.

	Recall of VISUAL Category		Recall of LEXICAL Category	
	Visual Input	Lexical Input (Identification)	Lexical Input	Visual Input (Description)
FA	37 (100%)	32 (86.5%)	32 (86.5%)	32 (86.5%)
FA (CC)	37 (100%)	37 (100%)	37 (100%)	37 (100%)
GA	37 (100%)	37 (100%)	37 (100%)	37 (100%)

Table 6.1: Performance of various ARTMAP models when trained for one epoch and then tested using Sales' training data. The columns of data show from left to right the results obtained for tests 1 through 4 (see main text). Each table cell shows how many patterns (out of a possible 37) correctly led to the the recall of the appropriate visual or lexical category.

whether recall of the correct lexical category was achieved.

4. Given a lexical input pattern in isolation each system was assessed to determine whether recall of the correct visual category was achieved (identification).

Tests 1 and 3 provide a means by which to assess how well each ART module has learnt the information with which it was provided. Tests 2 and 4 on the other hand act as an instrument through which one can determine how well each ARTMAP models Map Field has acquired associations between visual and lexical entities. Table 6.1 presents some of the findings that were obtained as a result of these testing procedures where fast learning was used for all ARTMAP weights (i.e. $\beta_a = \beta_b = \beta_{map} = 1.0$).

As one might expect, given that the five models within each group were identical and were trained and tested on identical data, the results obtained for each model within a group were exactly the same. All of the FACC, and GA models learnt the training data totally within one epoch and in each case 74 category units (37 visual $F2_a$ units & 37 lexical $F2_b$ units) were generated in order to encode the knowledge. As a result, in each of the four tests these systems registered 100% accuracy. Conversely, the FA models failed to correctly learn the training data. As reported in table 6.1, when recalling the data the FA systems failed to select all of the appropriate categories in three of the four tests. In each case five input patterns resulted in errors. Further analysis indicated that whilst 37 visual categories had been correctly created, only 32 lexical categories existed — a deficit of 5 categories.

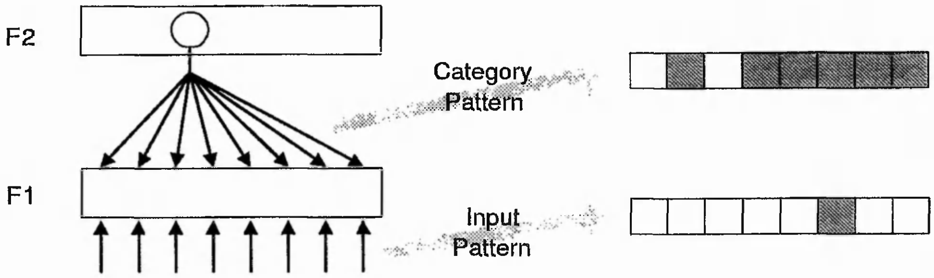

Figure 6.3: An input pattern maximally activates an F2 unit since it is a subset of the category pattern stored by the units weights. With fast learning the category pattern is updated so that it equals the input pattern.

These problems were traced to the internal pattern comparison mechanism used in FA models. In such models an input pattern can maximally activate a category unit under any conditions if it is a subset of the prototype category pattern that the unit has acquired. It matters not if vigilance is set to its maximum value, or if in vector space terms the input pattern is largely dissimilar to the category pattern (as shown in fig. 6.3).

The data set devised by Sales contains a number of patterns that are a subset of other patterns, for example on the visual side: chair $\subset$ door, daddy $\subset$ mummy, and spoon $\subset$ house (see figs. 6.1 & 6.2). On the lexical side: 'daddy' $\subset$ 'dolly', 'apple' $\subset$ 'cup', and 'bat' $\subset$ 'cat'. Hence, in the case of training as described above the following scenario unfolds[3]: The first 6 input pairs (cat-'cat', dog-'dog', clock-'clock', chair-'chair', dolly-'dolly', & mummy-'mummy') presented to the FA model are learnt correctly, i.e. the visual and lexical patterns are coded by ART$_a$ units $J = 0$ through $J = 5$ and ART$_b$ units $K = 0$ through $K = 5$, and the mapfield weights appropriately associate $J = 0 \Leftrightarrow K = 0$, $J = 1 \Leftrightarrow K = 1$ etc. Complications now begin as both patterns within the 7th input pair (daddy-'daddy') are a subset of patterns already acquired by the FA model — daddy $\subset$ mummy, and 'daddy' $\subset$ 'dolly'. As a result, the visual and lexical patterns respectively activate and select the category units $J = 5$ and $K = 4$. Because mapfield associations have already been forged that link $K = 4$ to $J = 4$, and $J = 5$ to $K = 5$, a mapfield mismatch is triggered. This resets unit $J = 5$ and selects a new ART$_a$ unit $J = 6$. Learning ensues whereby the lexical pattern for 'daddy' overwrites 'dolly' on $K = 4$, the daddy pattern is learnt by $J = 6$, and

[3] This particular scenario and the resulting problems arise because of the order in which the data was presented to the models during training. Similar problems would occur to a greater or lesser extent depending upon the presentation order.

a mapfield bond is forged between $J = 6$ and $K = 4$. Thus, there are now separate links between two ART_a units and a single ART_b unit. These multiple links can be seen in figure 6.4 which shows the final state of the network after one whole training epoch.

Subsequent training epochs don't necessarily redress this problem, in fact they can cause further problems. It is true that further training will remedy the deficit of ART_b units. However, as a result category proliferation can occur in the ART_a module with multiple units coding the same input pattern. In the case above after a second epoch of fast learn training, it was found that whilst 37 ART_b units had been generated as required (one for each input pattern), an excess of 6 ART_a units had also been generated giving 43 ART_a units in total. Six of the visual patterns (cat, dolly, cup, mouse, moon, & sun) had been coded by two separate units as shown in figure 6.5.

There *are* ways to enhance FA performance, however. One way is to bias the system to select units with a high index value. If after a number of training epochs there is a position where two ART_a units code an input pattern (e.g. dolly is coded by $J = 4$ and $J = 38$ in figure 6.5) it is likely that the unit with the highest index (e.g. $J = 38$) will correctly map to the appropriate ART_b unit. A second, equally valid approach would be to reduce the rate at which permanent mapfield associations become acquired (i.e. slow mapfield learning) thus reducing the likelihood of spurious connections. However, it should be noted that whichever method is preferred one still sacrifices autonomy to the extent that one must be aware of the data and tweak parameters of the model accordingly (i.e. $\beta_{map}\&\rho_{map}$). For the purposes of the investigations described throughout this chapter all of the FA models were trained over two whole epochs of the training data, and the bias method was employed to select high index units in the testing phase.

As predicted above each model was able to learn the training data set in a very short space of time. Table 6.2 shows the total times required to train each of the fifteen networks in question and the training time per individual pattern pair. In each case it can be seen that a pair of patterns is learnt in under 100 milliseconds. Given that the computing equipment used in these tests was far from being considered state-of-the-art, the results provide no evidence that constrain the practical application of an ARTMAP based model.

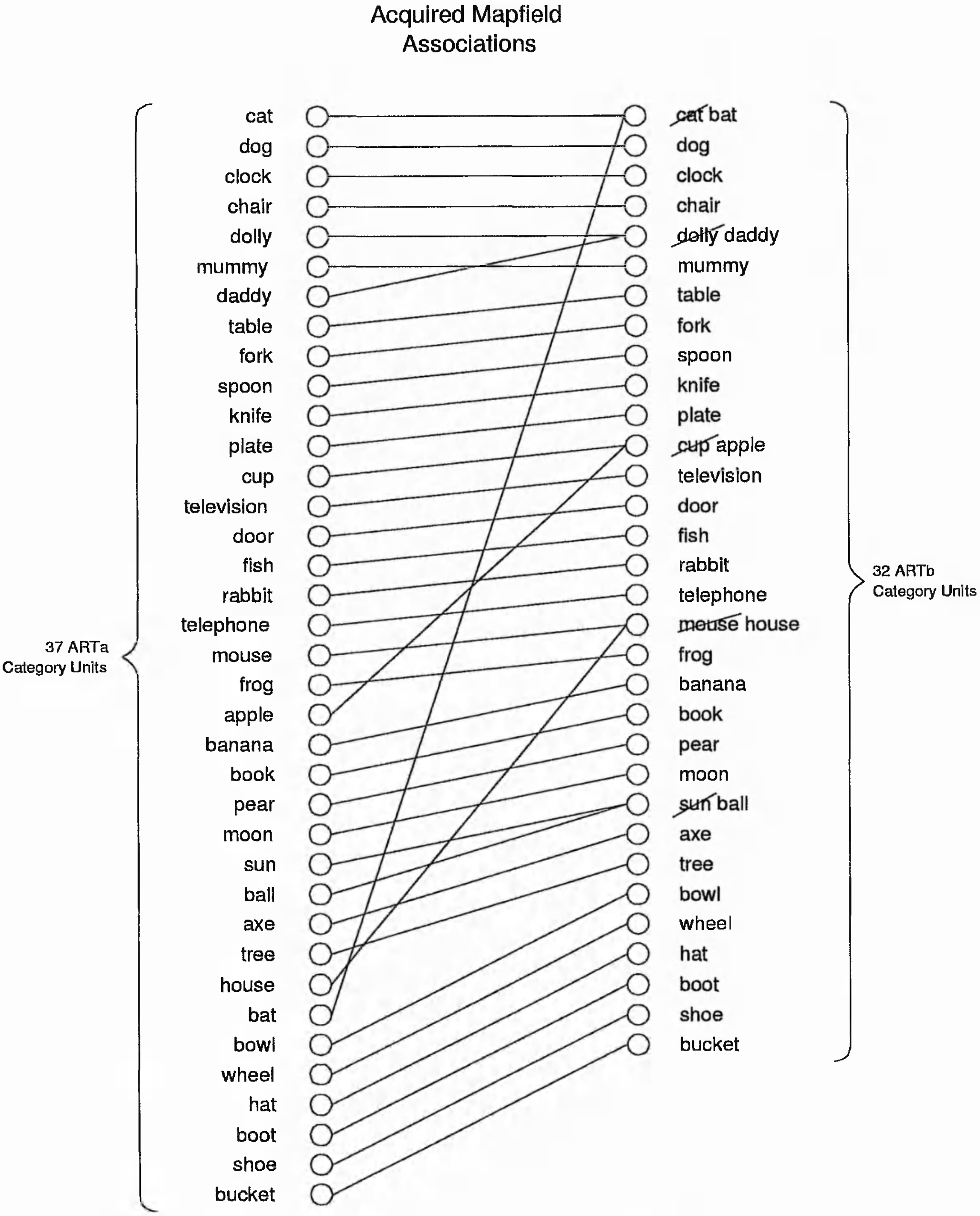

Figure 6.4: After one epoch of training the Fuzzy ARTMAP model has acquired 37 visual category units and 32 lexical category units. Because a number of ART$_b$ units have been recoded during learning the mapfield contains a number of spurious connections. For example, initially a link was forged between ART$_a$ unit 0 and ART$_b$ unit 0 (shown at top of figure) when the units respectively coded the visual cat and lexical 'cat' patterns. During the course of later learning 'bat' overwrote the category 'cat'. The model is left with a link that joins cat to 'bat'.

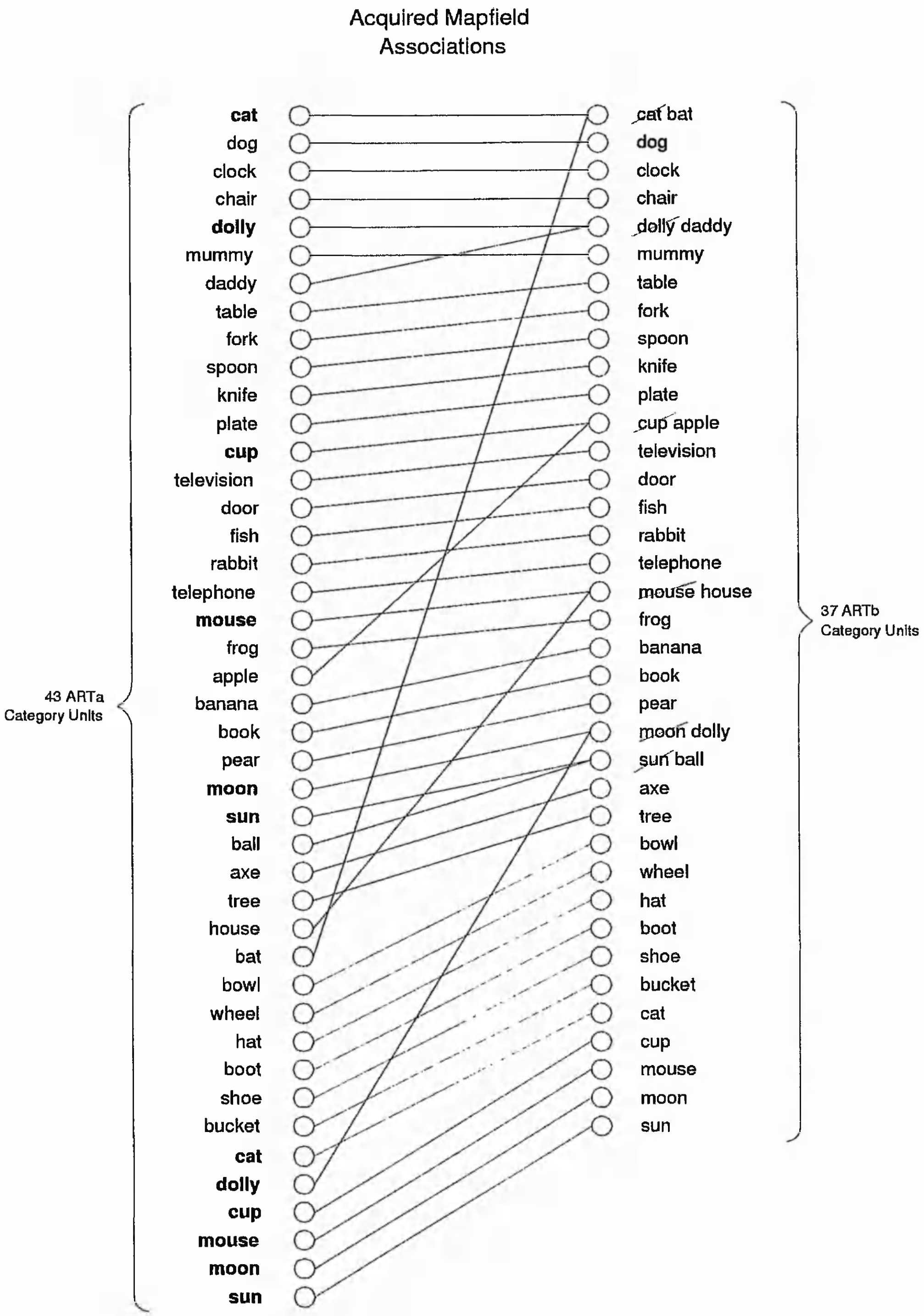

Figure 6.5: Shows state of Fuzzy ARTMAP mapfield after 2 epochs of training. patterns shown in bold have been coded by multiple ART category units

	FA training times		FACC training times		GA training times	
	total	per pattern	total	per pattern	total	per pattern
1	2.40	0.065	3.35	0.091	2.58	0.070
2	2.42	0.065	3.24	0.088	2.63	0.071
3	2.41	0.065	3.24	0.088	2.47	0.067
4	2.42	0.065	3.24	0.088	2.58	0.070
5	2.43	0.066	3.27	0.088	2.53	0.068
Av.	**2.42**	**0.065**	**3.25**	**0.088**	**2.56**	**0.069**

Table 6.2: Training times. Two columns of data are shown for each model. The left column shows the total time taken to train the model on all 37 patterns. The right column shows the training time required per pattern (i.e. left column divided by 37). The totals relate to two epochs of FA training, and 1 epoch of FACC and GA training. All times are shown in seconds.

6.1.3 Testing the Models

In order to further assess the fifteen ART based models introduced above, a body of testing data was generated. As per Sales' original investigations this testing data was created by adding noise to the training data set. Each pattern had 10% noise applied. This was acheived by reversing the value of 10% of the bits in the patterns, i.e. 14 of the 144 bits were reversed in the visual patterns, 6 of 64 were reversed in the linguistic patterns. All in all five different testing data sets were generated each consisting of 222 visual-linguistic pattern pairs (six noisy pattern pairs were generated from each of the 37 visual-linguistic pairs in the original training data).

Each model was trained as before (GA and FACC models for one epoch, FA models for two). However, this time rather than being tested on the same data (i.e. the training set) each model within an ARTMAP group was tested using one of the five generated testing pattern sets. The testing process again involved assessing the performance of knowledge recall from a particular model using the four tests described in the preceeding section. Tables 6.3, 6.4, and 6.5 detail the findings acquired from these investigations.

It is immediately clear from the first table of the trio that Fuzzy ARTMAP based models that do not employ complement coding are at a distinct disadvantage compared to the other

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
FA 1	167 (75.23%)	187 (84.23%)	187 (84.23%)	167 (75.23%)
FA 2	174 (78.38%)	190 (85.59%)	190 (85.59%)	174 (78.38%)
FA 3	174 (78.38%)	182 (81.98%)	182 (81.98%)	174 (78.38%)
FA 4	166 (74.77%)	188 (84.68%)	188 (84.68%)	166 (74.77%)
FA 5	164 (73.87%)	187 (84.23%)	187 (84.23%)	164 (73.87%)
Average	169 (76.13%)	186.8 (84.14%)	186.8 (84.14%)	169 (76.13%)

Table 6.3: Performance of Fuzzy ARTMAP models in relation to noisy test data. The table provides detail of how many correct patterns were recalled (out of 222 patterns) and shows the mean performance of all the networks taking into account all five noisy data sets.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
FACC 1	214 (96.40%)	222 (100.0%)	222 (100.0%)	214 (96.40%)
FACC 2	217 (97.75%)	222 (100.0%)	222 (100.0%)	217 (97.75%)
FACC 3	218 (98.20%)	222 (100.0%)	222 (100.0%)	218 (98.20%)
FACC 4	218 (98.20%)	222 (100.0%)	222 (100.0%)	218 (98.20%)
FACC 5	215 (96.85%)	222 (100.0%)	222 (100.0%)	215 (96.85%)
Average	216.4 (97.48%)	222 (100.0%)	222 (100.0%)	216.4 (97.48%)

Table 6.4: Performance of Fuzzy ARTMAP (employing Complement Coding) models in relation to noisy test data. The table provides detail of how many correct patterns were recalled (out of 222 patterns) and shows the mean performance of all the networks taking into account all five noisy data sets.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	214 (96.40%)	222 (100.0%)	222 (100.0%)	214 (96.40%)
GA 2	217 (97.75%)	222 (100.0%)	222 (100.0%)	217 (97.75%)
GA 3	218 (98.20%)	222 (100.0%)	222 (100.0%)	218 (98.20%)
GA 4	218 (98.20%)	222 (100.0%)	222 (100.0%)	218 (98.20%)
GA 5	215 (96.85%)	222 (100.0%)	222 (100.0%)	215 (96.85%)
Average	216.4 (97.48%)	222 (100.0%)	222 (100.0%)	216.4 (97.48%)

Table 6.5: Performance of Gaussian ARTMAP models in relation to noisy test data. The table provides detail of how many correct patterns were recalled (out of 222 patterns) and shows the mean performance of all the networks taking into account all five noisy data sets.

ARTMAP types investigated. Averaged over five result sets the FA solution was able to recall only 169 visual patterns correctly from a total of 222 lexical patterns (76% success) and only 187 lexical patterns from the 222 visual patterns (84% success). An interesting point to note is the commonality that is present in the results with columns 1 and 4, and columns 2 and 3 of table 6.3 containing identical data. This suggests that the data learnt in the models intra-module mapfield (the semantic knowledge) is largely correct and that all the predictive failures within the system are arising from recall failures within each ART module, i.e. a visual pattern of a dog is not being visually recognised as a dog.

Tables 6.3, 6.4 and 6.5 show that the results acheived using the FACC and GA models far outstrip those acheived using plain FA models. For FACC and GA models, visual patterns were visually recognised correctly 100% of the time and also in each case the correct counterpart lexical label was recalled correctly without any need for lexical input. Whilst performance with lexical input alone was not quite as successful, it was still more than satisfactory with all of the FACC and GA models acheiving over 96% success in recall of both the correct lexical category and then by association the correct visual category. Again the symmetry in data that can be seen in both tables points to the fact that recognition failures were occuring within the ART modules, not in the intra-module mapfield. These failures are considered further in the following section.

	FA test times		FACC test times		GA test times	
	total	per pattern	total	per pattern	total	per pattern
1	3.19	0.014	4.89	0.022	4.07	0.018
2	3.24	0.015	4.61	0.021	3.96	0.018
3	3.24	0.015	4.45	0.020	4.01	0.018
4	3.30	0.015	4.51	0.020	4.18	0.019
5	3.29	0.015	4.45	0.020	3.96	0.018
Av.	**3.25**	**0.015**	**4.58**	**0.021**	**4.04**	**0.018**

Table 6.6: Operating times. As in table 6.2 two colums of data are provided for each type of model. This time the left column shows the cummulative time taken by the model to make predictions for all 222 test patterns. The right side column shows the time for recall of knowledge required per pattern. All times are shown in seconds.

In addition to obtaining the success/failure in prediction data discussed above the operating times required by each model to make a prediction based upon an unknown noisy input was measured. Table 6.6 details the results of these measurements. In line with the training times detailed above it can be seen that the operational times that result from using ART models are not really at issue. In *all* the test cases these models were able to make a prediction in less than a quarter of a second. The FACC models consistently required more time than the other models. This is largely due to complement coding in that many more vector calculations are required to process the larger input patterns.

6.1.4 Introducing More Noise

The results obtained in the experiments above adequately demonstate that ART models are quite capable of handling noisy data to the extent investigated by Sales, i.e. at the 10% level. In order to fully assess ART's potential this section describes futher experimentation, results, and analysis.

For these further investigations a new body of testing data was generated. This body contained 10 groups of noisy data, each containing 5 different testing data sets comprised of 222 visual-linguistic pattern pairs. The 10 groups had noise levels varying in 2% increments from 12% noise (i.e. 17 of 144 visual bits reversed & 8 of 64 linguistic bits reversed) through to 30% noise (43 visual & 19 linguistic bits reversed). All of the tests were carried out using Gaussian ARTMAP systems only. The rationale for this focus being Gaussian ARTMAP's overall abilities in both acuracy and speed of knowledge recall as demonstrated above.

Each model was trained as previously and then tested using one of the new noisy data sets. Thus, all in all a total of fifty (50) Gaussian ARTMAP models were trained and tested. Table 6.7 details the results obtained from the testing process. Each line of the table provides the figures averaged from 5 GAM models.

As one might expect it is immediately clear from the results that as more and more noise is added to the data, performance in knowledge recall deteriorates. Also evident is the fact that this deterioration is greater for the recall of knowledge from lexical input data alone (columns 1 & 4 in the table) than it is when visual input data is used in isolation. In the 30% noise tests it can be seen that this divide in performance has reached a level of nearly 40% — a significant factor.

The reasons for these differences can be explained with the aid of further analysis of Sales' original data (see figs. 6.1 & 6.2). Hamming distances, d$_H$, are a measure between points within an I dimensional binary space. They provides a means by which to judge the similarity between two binary vectors of equal length by comparing bits thus:

$$d_H(\overline{x},\overline{y}) = \sum_{i=1}^{I} |x_i - y_i| \tag{6.1}$$

Therefore, considering the lexical vectors for the words 'book' and 'dog':

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
12%	209.4 (94.32%)	222.0 (100.0%)	222.0 (100.0%)	209.4 (94.32%)
14%	204.6 (92.16%)	222.0 (100.0%)	222.0 (100.0%)	204.6 (92.16%)
16%	205.6 (92.61%)	221.8 (99.91%)	221.8 (99.91%)	205.6 (92.61%)
18%	189.4 (85.32%)	221.6 (99.82%)	221.6 (99.82%)	189.4 (85.32%)
20%	186.4 (83.96%)	221.2 (99.64%)	221.2 (99.64%)	186.4 (83.96%)
22%	179.4 (80.81%)	221.0 (99.55%)	221.0 (99.55%)	179.4 (80.81%)
24%	173.6 (78.20%)	220.0 (99.10%)	220.0 (99.10%)	173.6 (78.20%)
26%	155.6 (71.10%)	220.0 (99.10%)	220.0 (99.10%)	155.6 (71.10%)
28%	141.2 (63.60%)	218.4 (98.38%)	218.4 (98.38%)	141.2 (63.60%)
30%	128.4 (57.84%)	214.0 (96.40%)	214.0 (96.40%)	128.4 (57.84%)

Table 6.7: Performance of Gaussian ARTMAP models on noisy test data.

$$book = \{\ 0001001111011110101111111111111111111111111111111111111111111111\ \}$$

$$dog = \{\ 0010001111001111111111111111111111111111111111111111111111111111\ \}$$

$$d_H(\overline{book}\,,\,\overline{dog}\,) = 5$$

The smaller the Hamming distance the greater the similarity between two vectors and the greater the probability that the addition of noise will lead to recall errors. Consider the following two noisy 'book' patterns:

$$book_1 = \{\ 0101001111111110101101111111110111111111011111111101111111111111\ \}$$

$$book_2 = \{\ 0010001111011111101111111111111011111111111111111111011111110111\ \}$$

each of which have had six bits randomly reversed by noise. Bits 1,10,20,30,40 & 50 have been reversed in pattern 1, bits 2,3,15,31,52 & 60 in pattern 2. The application of equation 6.1 provides the following Hamming distances:

CASE 1:

$$d_H(\overline{book_1}\,,\,\overline{book}\,) \quad = \quad 6$$

$$d_H(\overline{book_1}\,,\,\overline{dog}\,) \;=\; 11$$

CASE 2:

$$\begin{aligned} d_H(\overline{book_2}\,,\,\overline{book}\,) &= 6 \\ d_H(\overline{book_2}\,,\,\overline{dog}\,) &= 5 \end{aligned}$$

The application of noise in case 1 has has resulted in a pattern *book*$_1$ that still shares greater similarity with its originating pattern *book* than it does with the alternate pattern *dog*. However, in the second case the noise addition has resulted in a pattern *book*$_2$ that now shares more in common with *dog* than it does with *book*. In this case one would expect from a correctly functioning ARTMAP model to suggest that *book*$_2$ is in fact a *dog*.

The probability of a failure occurring as above is therefore a function of (1) the hamming distance between patterns, and (2) the amount of noise added to a pattern.

Figures 6.6 and 6.7 provide a detailed picture of all of the Hamming distances between the lexical input and visual input patterns respectively. The figures show clearly that the Hamming distances between the lexical patterns are less than those between the visual patterns. Further calculations show that the average Hamming distances between the lexical and visual patterns are 12.8 and 46.4 bits respectively. Thus the average Hamming distance within a lexical pattern is 20% of the patterns overall length. For a visual pattern this figure rises to 32%. It is therefore obvious that given a fixed percentage addition of noise the recall of lexical knowledge will be poorer than for visual knowledge.

To conclude an analysis of the results it should be noted that as before there is a clear symmetry within the results and furthermore that this is consistent with the fact that the lexical knowledge acquired by the models during training is accurate. The failures can again be attributed to problems of recall within each ART module (as discussed in this section).

	cat	dog	clock	chair	dolly	mummy	daddy	table	fork	spoon	knife	plate	cup	television	door	fish	rabbit	telephone	mouse	frog	apple	banana	book	pear	moon	sun	ball	axe	tree	house	bat	bowl	wheel	hat	boot	shoe	bucket
cat		9	12	11	13	12	10	13	9	10	15	15	3	23	13	10	15	24	12	11	10	14	10	10	12	6	6	6	12	12	1	9	12	3	11	10	15
dog	9		11	14	8	11	11	12	6	13	14	12	10	22	4	9	18	23	9	8	17	17	5	9	5	9	11	7	9	9	8	6	13	8	6	11	14
clock	12	11		9	13	12	14	17	11	8	9	13	13	25	9	12	13	24	12	9	16	12	8	12	10	10	12	10	14	12	13	11	10	15	9	6	13
chair	11	14	9		14	13	13	14	12	11	12	12	12	24	16	9	16	23	15	14	11	15	11	13	15	15	11	9	13	15	12	12	9	14	14	9	14
dolly	13	8	13	14		7	5	10	10	17	12	12	14	20	8	11	16	17	13	12	13	17	9	13	7	13	7	13	11	13	12	8	13	12	8	13	18
mummy	12	11	12	13	7		8	13	15	12	13	15	11	21	13	14	19	18	12	11	12	18	12	14	8	10	10	10	10	14	13	13	12	11	13	12	15
daddy	10	11	14	13	5	8		7	13	16	15	11	13	19	11	10	11	16	14	11	10	14	14	10	12	14	8	12	8	14	9	11	10	9	11	12	17
table	13	12	17	14	10	13	7		12	13	14	8	14	20	14	9	10	19	15	14	7	15	15	11	13	13	9	15	11	15	12	12	11	12	14	13	12
fork	9	6	11	12	10	15	13	12		15	14	16	8	24	8	5	14	25	11	10	15	15	5	11	9	11	11	13	13	11	8	6	15	10	10	13	12
spoon	10	13	8	11	17	12	16	13	15		13	15	11	23	15	14	15	24	18	7	12	12	10	14	12	6	10	8	10	18	11	13	8	13	13	6	13
knife	15	14	9	12	12	13	15	14	14	13		8	14	26	12	15	16	25	9	12	11	19	11	13	9	15	13	13	15	9	16	12	11	16	10	11	14
plate	15	12	13	12	12	15	11	8	16	15	8		14	22	12	13	14	21	11	16	9	19	15	7	15	15	13	11	11	9	14	12	9	14	10	11	14
cup	3	10	13	12	14	11	13	14	8	11	14	14		24	14	11	16	25	13	12	9	17	11	9	13	5	9	7	13	13	4	10	15	6	12	13	12
television	23	22	25	24	20	21	19	20	24	23	26	22	24		22	25	26	13	25	22	25	17	23	21	23	21	19	25	19	25	22	24	21	22	22	21	26
door	13	4	9	16	8	13	11	14	8	15	12	12	14	22		9	16	23	9	8	19	15	5	7	5	11	11	11	11	9	12	8	15	12	4	11	16
fish	10	9	12	9	11	14	10	9	5	14	15	13	11	25	9		13	26	14	11	12	14	8	10	10	14	8	12	12	14	9	5	12	11	9	10	13
rabbit	15	18	13	16	16	19	11	10	14	15	16	14	16	26	16	13		21	19	16	15	11	15	11	19	15	13	19	17	17	14	16	15	16	16	15	8
telephone	24	23	24	23	17	18	16	19	25	24	25	21	25	13	23	26	21		24	23	24	18	24	22	24	22	20	26	20	24	23	25	20	23	23	22	23
mouse	12	9	12	15	13	12	14	15	11	18	9	11	13	25	9	14	19	24		15	14	20	12	12	10	16	18	12	14	2	13	13	12	11	13	16	15
frog	11	8	9	14	12	11	11	14	10	7	12	16	12	22	8	11	16	23	15		15	13	7	13	9	9	9	9	5	15	10	10	11	12	8	7	14
apple	10	17	16	11	13	12	10	7	15	12	11	9	9	25	19	12	15	24	14	15		18	18	14	16	14	10	10	12	14	11	13	12	11	17	14	13
banana	14	17	12	15	17	18	14	15	15	12	19	19	17	17	15	14	11	18	20	13	18		12	16	18	14	10	18	16	18	13	15	14	15	13	12	13
book	10	5	6	11	9	12	14	15	5	10	11	15	11	23	5	8	15	24	12	7	18	12		10	6	8	8	10	12	10	9	5	14	11	5	8	11
pear	10	9	12	13	13	14	10	11	11	14	13	7	9	21	7	10	11	22	12	13	14	16	10		12	10	10	10	10	10	9	11	14	9	9	12	13
moon	12	5	10	15	7	8	12	13	9	12	9	15	13	23	5	10	19	24	10	9	16	18	6	12		10	10	10	12	12	13	7	14	11	7	10	17
sun	6	9	10	15	13	10	14	13	11	6	15	15	5	21	11	14	15	22	16	9	14	14	8	10	10		8	8	12	16	7	11	14	9	9	8	13
ball	6	11	12	11	7	10	6	9	11	10	13	13	9	19	11	8	13	20	18	9	10	10	8	10	10	8		10	10	16	5	7	12	7	7	8	15
axe	6	7	10	9	13	10	12	15	13	8	13	11	7	25	11	12	19	26	12	9	10	18	10	10	10	8	10		8	12	7	11	10	7	11	8	15
tree	12	9	14	13	11	10	8	11	13	10	15	11	13	19	11	12	17	20	14	5	12	16	12	10	12	12	10	8		14	11	11	8	11	11	8	17
house	12	9	12	15	13	14	14	15	11	18	9	9	13	25	9	14	17	24	2	15	14	18	10	10	12	16	16	12	14		11	11	14	9	11	16	13
bat	1	8	13	12	12	13	9	12	8	11	16	14	4	22	12	9	14	23	13	10	11	13	9	9	13	7	5	7	11	11		8	13	2	10	11	14
bowl	9	6	11	12	8	13	11	12	6	13	12	12	10	24	8	5	16	25	13	10	13	15	5	11	7	11	7	11	11	11	8		13	10	4	9	14
wheel	12	13	10	9	13	12	10	11	15	8	11	9	15	21	15	12	15	20	12	11	12	14	14	14	14	14	12	10	8	14	13	13		15	13	6	17
hat	3	8	15	14	12	11	9	12	10	13	16	14	6	22	12	11	16	23	11	12	11	15	11	9	11	9	7	7	11	9	2	10	15		12	13	16
boot	11	6	9	14	6	13	11	14	10	13	10	10	12	22	4	9	16	23	13	8	17	13	5	9	7	9	7	11	11	11	10	4	13	12		7	16
shoe	10	11	6	9	13	12	12	13	13	6	11	11	13	21	11	10	15	22	16	7	14	12	8	12	10	8	8	8	8	16	11	9	6	13	7		17
bucket	15	14	13	14	18	15	17	12	12	13	14	14	12	26	16	13	8	23	15	14	13	13	11	13	17	13	15	15	17	13	14	14	17	16	16	17	

Figure 6.6: The table above shows the Hamming distances (see main text for details) beweeen lexical patterns taken from Sales' data.

	cat	dog	clock	chair	dolly	mummy	daddy	table	fork	spoon	knife	plate	cup	television	door	fish	rabbit	telephone	mouse	frog	apple	banana	book	pear	moon	sun	ball	axe	tree	house	bat	bowl	wheel	hat	boot	shoe	bucket
cat	■	46	66	43	44	43	43	45	58	61	66	43	37	64	45	49	37	61	55	30	39	56	36	45	50	48	45	52	62	39	53	31	58	43	57	44	35
dog	46	■	74	57	54	49	53	51	64	71	78	45	45	74	57	43	33	53	61	46	43	66	50	55	58	52	47	68	66	49	67	41	62	51	71	54	43
clock	66	74	■	53	58	71	65	51	48	47	46	65	63	54	63	61	55	73	59	60	73	48	74	65	56	62	75	60	66	73	53	65	42	53	67	56	75
chair	43	57	53	■	37	48	36	42	39	36	43	32	44	65	26	52	52	60	58	45	42	37	35	32	47	43	44	37	63	48	40	40	39	36	50	43	52
dolly	44	54	58	37	■	33	21	43	36	29	36	31	45	72	27	37	47	63	63	50	39	34	36	33	36	36	35	38	42	39	27	39	48	37	63	50	43
mummy	43	49	71	48	33	■	12	42	43	44	51	24	38	79	26	42	44	60	64	47	34	51	23	32	47	45	20	47	39	8	40	40	59	48	44	57	32
daddy	43	53	65	36	21	12	■	46	31	32	39	20	36	79	14	36	46	66	68	47	30	39	21	28	43	37	20	37	33	20	28	32	53	40	50	57	32
table	45	51	51	42	43	42	46	■	41	38	41	38	38	63	52	36	50	64	62	39	36	35	39	36	43	35	50	41	63	50	42	34	43	14	54	41	42
fork	58	64	48	39	36	43	31	41	■	17	22	43	45	64	31	49	57	77	55	50	41	20	40	31	50	42	49	52	54	51	19	43	38	35	51	48	51
spoon	61	71	47	36	29	44	32	38	17	■	7	40	52	61	34	48	62	72	62	53	46	15	43	28	41	41	48	45	47	52	20	48	37	32	58	49	56
knife	66	78	46	43	36	51	39	41	22	7	■	47	53	60	41	51	63	79	59	54	47	18	48	29	46	42	55	50	50	59	17	49	36	35	53	46	59
plate	43	45	65	32	31	24	20	38	43	40	47	■	46	75	30	36	48	62	78	47	36	43	27	40	47	39	24	33	35	32	48	32	49	32	50	55	36
cup	37	45	63	44	45	38	36	38	45	52	53	46	■	61	38	32	34	58	36	35	18	45	21	30	41	37	34	53	63	38	40	20	55	36	36	35	24
television	64	74	54	65	72	79	79	63	64	61	60	75	61	■	83	77	77	47	53	60	67	60	62	75	58	82	69	58	72	79	69	73	58	67	59	58	69
door	45	57	63	26	27	26	14	52	31	34	41	30	38	83	■	48	48	66	58	51	36	39	21	26	47	39	30	37	41	22	30	42	57	42	48	57	38
fish	49	43	61	52	37	42	36	36	49	48	51	36	32	77	48	■	38	66	60	45	26	45	39	34	39	31	30	55	57	50	44	22	51	28	60	39	28
rabbit	37	33	55	52	47	44	46	50	57	62	63	48	34	77	48	38	■	64	50	31	36	57	43	46	49	43	44	67	63	42	54	30	51	48	62	41	40
telephone	61	53	73	60	63	60	66	64	77	72	79	62	58	47	66	66	64	■	60	57	62	69	65	68	43	69	50	63	59	58	74	64	61	66	66	61	50
mouse	55	61	59	58	63	64	68	62	55	62	59	78	36	53	58	60	50	60	■	49	48	55	53	44	49	63	66	71	91	60	54	54	63	60	42	27	58
frog	30	46	60	45	50	47	47	39	50	53	54	47	35	60	51	45	31	57	49	■	31	48	44	39	48	50	47	64	72	47	49	31	52	37	57	34	39
apple	39	43	73	42	39	34	30	36	41	46	47	36	18	67	36	26	36	62	48	31	■	41	23	24	47	35	26	53	61	36	36	16	43	32	40	31	22
banana	56	66	48	37	34	51	39	35	20	15	18	43	45	60	39	45	57	69	55	48	41	■	42	29	36	40	51	44	56	59	19	41	36	29	57	40	51
book	36	50	74	35	36	23	21	39	40	43	48	27	21	62	21	39	43	65	53	44	23	42	■	25	44	36	23	32	50	23	39	31	60	37	33	48	23
pear	45	55	65	32	33	32	28	36	31	28	29	40	30	75	26	34	46	68	44	39	24	29	25	■	39	29	32	51	57	36	18	30	47	30	44	33	30
moon	50	58	56	47	36	47	43	43	50	41	46	47	41	58	47	39	49	43	49	48	47	36	44	39	■	52	39	40	50	49	43	41	50	41	67	40	39
sun	48	52	62	43	36	45	37	35	42	41	42	39	37	82	39	31	43	69	63	50	35	40	36	29	52	■	47	52	60	53	35	33	52	29	57	46	41
ball	45	47	75	44	35	20	20	50	49	48	55	24	34	69	30	30	44	50	66	47	26	51	23	32	39	47	■	47	37	20	48	28	49	44	48	51	16
axe	52	68	60	37	38	47	37	41	52	45	50	33	53	58	37	55	67	63	71	64	53	44	32	51	40	52	47	■	48	49	51	49	56	39	51	62	51
tree	62	66	66	63	42	39	33	63	54	47	50	35	63	72	41	57	63	59	91	72	61	56	50	57	50	60	37	48	■	41	49	55	62	59	71	80	47
house	39	49	73	48	39	8	20	50	51	52	59	32	38	79	22	50	42	58	60	47	36	59	23	36	49	53	20	49	41	■	48	40	63	56	44	61	28
bat	53	67	53	40	27	40	28	42	19	20	17	48	40	69	30	44	54	74	54	49	36	19	39	18	43	35	48	51	49	48	■	40	43	36	54	45	48
bowl	31	41	65	40	39	40	32	34	43	48	49	32	20	73	42	22	30	64	54	31	16	41	31	30	41	33	28	49	55	40	40	■	39	28	48	33	20
wheel	58	62	42	39	48	59	53	43	38	37	36	49	55	58	57	51	51	61	63	52	43	36	60	47	50	52	49	56	62	63	43	39	■	43	61	42	51
hat	43	51	53	36	37	48	40	14	35	32	35	32	36	67	42	28	48	66	60	37	32	29	37	30	41	29	44	39	59	56	36	28	43	■	58	37	40
boot	57	71	67	50	63	44	50	54	51	58	53	50	36	59	48	60	62	66	42	57	40	57	33	44	67	57	48	51	71	44	54	48	61	58	■	43	46
shoe	44	54	56	43	50	57	57	41	48	49	46	55	35	58	57	39	41	61	27	34	31	40	48	33	40	46	51	62	80	61	45	33	42	37	43	■	47
bucket	35	43	75	52	43	32	32	42	51	56	59	36	24	69	38	28	40	50	56	39	22	51	23	30	39	41	16	51	47	28	48	20	51	40	46	47	■

Figure 6.7: The table above shows the Hamming distances between visual patterns taken from Sales' data.

6.1.5 An empirical comparison of ART based systems with WANNs

With the experimentation and results from the sections above now in place this section returns to the primary aim of the chapter, i.e. to compare the performance of ARTMAP against WANN based systems on a lexical acquisition and grounding task. The comparisons consider three empirical quantities:

1. The time required for a system to record information, i.e. training time.
2. The time required for a system to recall information, i.e. operating time.
3. The accuracy of the information recalled.

Training times

> "Typical training times for the networks were 30-40 seconds — around 1 second per training pattern learned" (Sales 1996, p. 186)

It is unquestionable from the data presented in figure 6.8 that ARTMAP systems have the upper hand when it comes to training speed. All of the ARTMAP models employed in the investigations above were faster than WANNs at learning data by roughly a factor of 10. A fairly significant advantage.

Operating times

On the topic of operating times Sales' remains a little vague. It is noted that in the WANN model operating times are a function the networks size (or in Sales terminology – storage cost). Sales provides operating time figures for only 2 of the 30 network variations that were employed. Networks 1 and 15 — the smallest and largest of the networks — required 3.9 and 16.2 seconds respectively to make a prediction per pattern (Sales 1996, p. 186). Based on an assumption that operating time rises linearly in relation to network size one can assume from the above figures that the networks that performed best in Sales' investigations (see next section) would have required in the order of 6 to 8 seconds to perform recall.

Figure 6.9 shows the operating times achieved by Sales employing WANNs and those

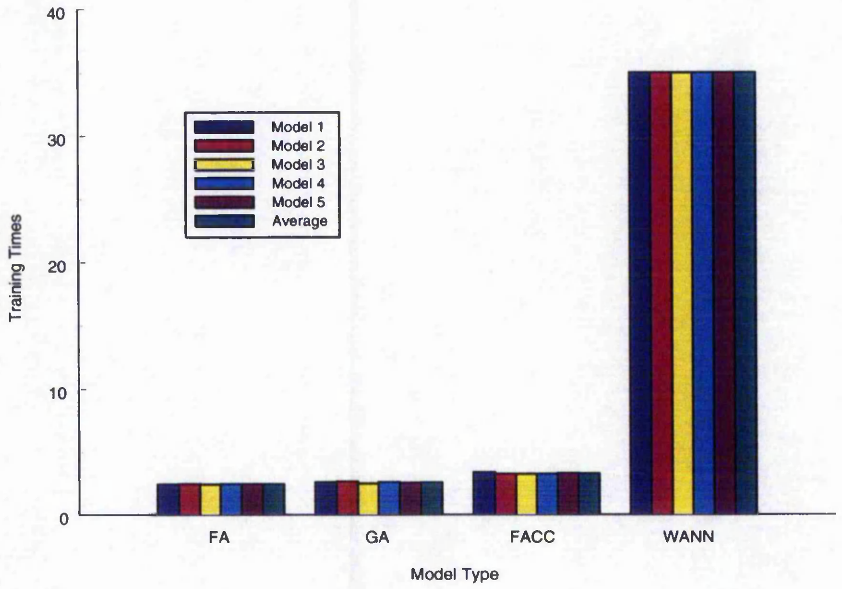

Figure 6.8: The figure shows the time required for the five models of each architecture type to learn all 37 patterns from the training data set (see also figure 6.2).

achieved in the investigations detailed above (see table 6.6). In the case of operating times the divide in performance between WANNs and ART based systems is even greater than those for training times already shown above. In fact the divide is so great that in the figure, a visualisation of the operating times for the ART based models is lost because of the scale necessary to plot the WANN results. With recall times in the order of 20 milliseconds per pattern all of the ART systems operate over 100 times faster than the fastest of the WANN models and about 300 to 400 times faster than the WANN models that performed the LAG task best. This is a very significant performance advantage. In a human-machine communication environment operating times of WANNs are simply unfeasable. Having issued a natural language command one would not expect to have to wait 6 to 8 seconds for the command to be actioned.

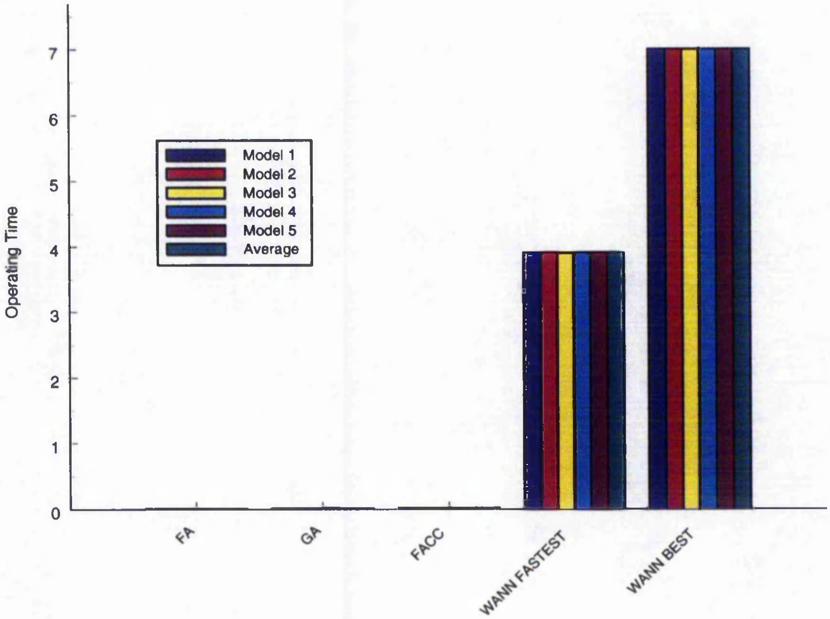

Figure 6.9: The figure shows the time required (in seconds) for the five models of each architecture type to recall a learnt pattern based on an input pattern (see also figure 6.6).

Accuracy of recall

As already noted in chapter 3 Sales employed a total of 90 WANN networks in his investigations that fell into three groupings:

1. 45 networks with balanced connectivity

2. 24 networks with connectivity biased to linguistic processing

3. 21 networks with connectivity biased to visual processing

Figures 6.10 to 6.13 detail the accuracy of recall for the four tests identified at the beginning of the chapter. In each figure the results from 3 WANN models (the best performing WANN model from each of the groupings discussed above) are plotted against the average results from all the tests detailed in the sections above (see tables 6.3 to 6.5 & 6.7). Each graph is now discussed in detail.

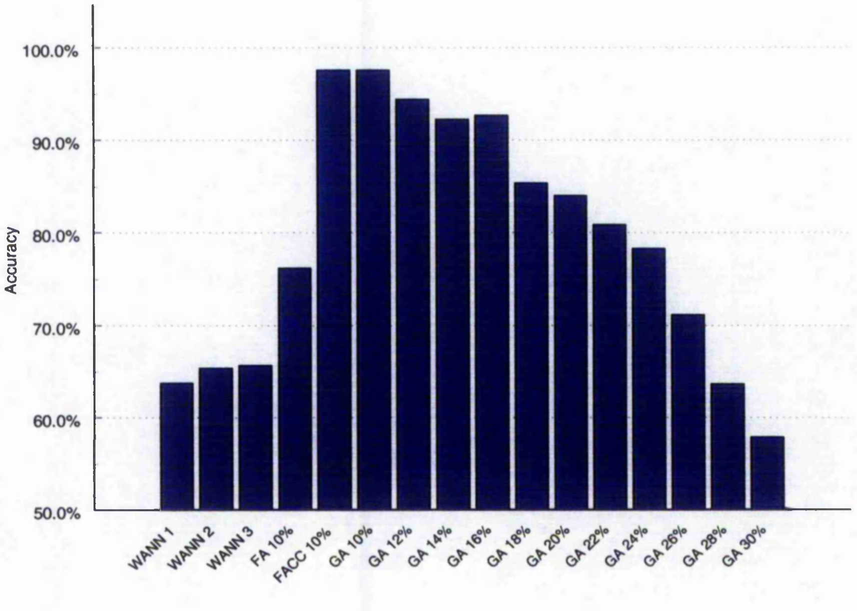

Figure 6.10: Graph showing performance on lexical → lexical task. Plotting % accuracy for 3 WANN networks vs. FA, FACC, & GA on 10% noisy data and also 10x GA (12% - 30% noise)

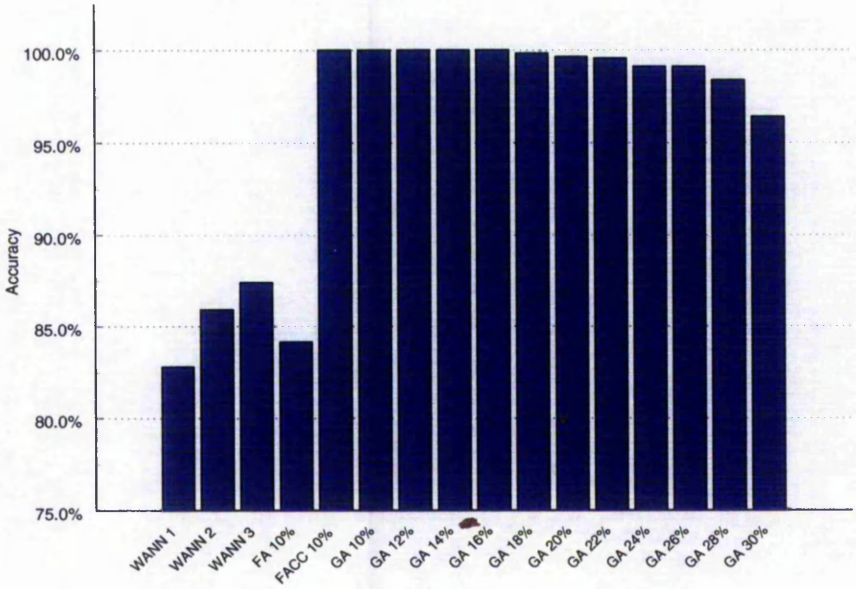

Figure 6.11: Graph showing performance on visual → lexical task. Plotting % accuracy for 3 WANN networks vs. FA, FACC, & GA on 10% noisy data and also 10x GA (12% - 30% noise)

It has already been discussed above that due the nature of the data, ART based systems will tend to fair better recalling past learning from visual input rather than from that which is lexical. Despite this fact figure 6.10 shows the ART based models tested using equivalent 10% noisy data all outperforming each of the WANN models on a lexical input / lexical recall task. Taking this further the majority of the GA models tested on data with a greater noise component than that employed by Sales also manage to outperform each of the WANN models on this task. In fact it is only when one gets to those models to the far right of the figure – those tested on data with greater than 26% added noise – that one sees performance figures in line and below those achieved by Sales' best performing networks. Again this is quite a significant result.

Moving to a description task (visual input / lexical recall) figure 6.11 again shows the clear performance superiority that most of the ART models have over their WANN counterparts — even in the presence of far greater noise levels. This same pattern of superiority is

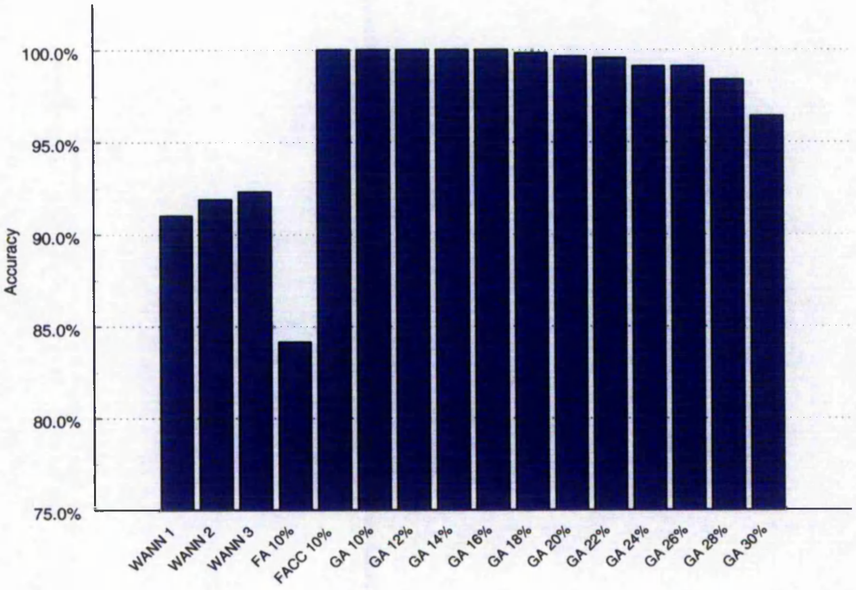

Figure 6.12: Graph showing performance on visual → visual task. Plotting % accuracy for 3 WANN networks vs. FA, FACC, & GA on 10% noisy data and also 10x GA (12% - 30% noise)

also evident in figure 6.12 which shows how each of the networks faired in a visual in / visual out task. It can be seen in both of these graphs that the FA type ART models are the only models that under-perform the WANN models tested. One can put this down to the problem of visual input patterns being subsets of other patterns as discussed above in section 6.1.2.

Finally, figure 6.13 shows the data acquired during the identification tests (lexical in / visual out). This final graph again provides evidence that an ARTMAP type architecture is better suited to a lexical acquisition and grounding task than a WANN architecture. In this case all but one of the ARTMAP models achieved results ranking above the competition.

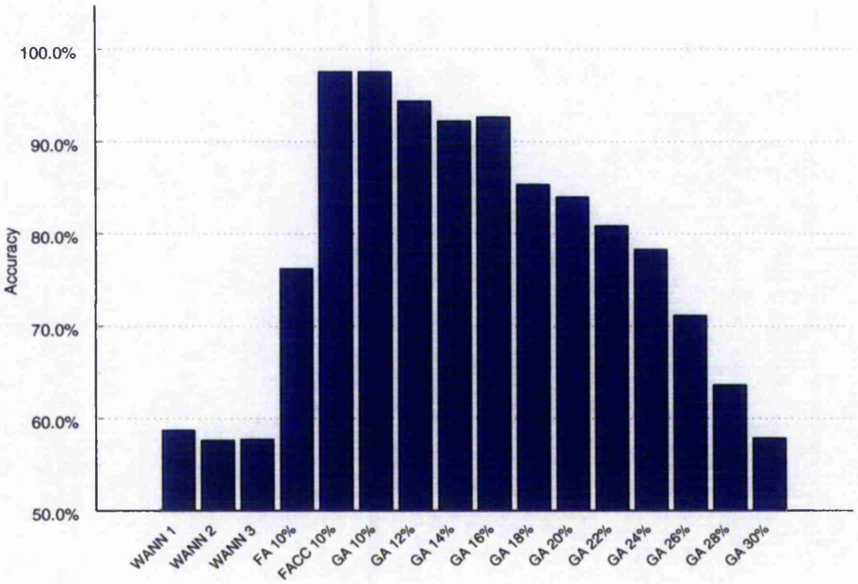

Figure 6.13: Graph showing performance on lexical → visual task. Plotting % accuracy for 3 WANN networks vs. FA, FACC, & GA on 10% noisy data and also 10x GA (12% - 30% noise)

6.1.6 ART vs. WANNS: A Final Analysis

The primary objective of this chapter was to substantiate the claim that ART-based lexical acquisition and grounding architectures are superior in many ways to other contemporary architectures. The investigations detailed above provided scope for a direct comparison between ART-based architectures and WANN architectures and thus a means by which to substantiate the above claim. And in all cases the results obtained from these investigations demonstrate clearly that in terms of three quantifiable measures ART systems are indeed superior to WANNs. They excel in their ability to learn about data quickly and then recall this data both quickly and accurately.

Sales' work focussed strongly upon the issue of "applicability to practical usage in realistic environments." (Sales 1996, p.178). To bring this chapter to a close this issue is considered again. It is used below to provide scope for further direct comparisons between ARTMAP and WANN architectures.

In WANN architectures both generalisation and the topology of the networks (number of nodes, weight connectivity etc.) are set explicitly based upon the designers a priori knowledge of the environment. In both cases these settings can impact significantly upon performance (for evidence, see (Sales 1996, pp 223-225)). In ARTMAP systems this is not the case, for example, the generalisation mechanism is itself adaptable and is directed by the environment itself without human intervention. In a realistic learning environment a designer would be fortunate to know everything about the environment in advance. Even in the case where it may be possible, it is also possible that the designer could overlook things. Thus, in this respect the benefits of the autonomous operation offered by ART based systems speaks for itself.

In discussing WANNs and the VRAM nodes upon which they are based Sales' provides the following admissions:

> "The VRAM is a particularly good solution for problems where a known number of input patterns ... need to be learned. The VRAM is not suitable for use where an upper bound for the number of training patterns is not known in advance." (Sales 1996, p.61)

Again in a realistic learning environment it cannot be assumed that one will know the number of objects, colours, words etc. that a system may be exposed to. This places a constraint on WANN systems that one need not worry about when using systems based instead on ART.

One final point worth addressing again is that WANNs employ a learning mechanism that functions from a basis of prototype data. In a realistic learning environment one is provided only with the noisy real world and not with prototype data. By contrast as the investigations in the previous chapter demonstrated ARTMAP systems can form their own prototypes based upon noisy data.

Chapter 7

A Multi-Modal Acquisition and Grounding Architecture

The previous two chapters have considered how connectionist architectures based upon Adaptive Resonance Theory can be used in various lexical acquisition and grounding tasks. These architectures have been shown to provide distinct performance benefits when compared to many of the CMs discussed in chapter 3. The final stage of the work presented in this thesis embraces and extends these investigations by proposing a learning model based upon ART principles that can acquire and ground diverse lexical terms in the presence of multiple sensory stimuli. This task is referred to as multi-modal (MM) grounding.

In this chapter the task of MM grounding is considered in some detail. A number of issues are addressed that question the applicability of a standard ARTMAP model to this task. On this basis, a new learning architecture based on ARTMAP is introduced, discussed, and then explored empirically *vis-à-vis* a practical MM grounding task, namely that of acquiring and grounding a lexicon that relates to both colour and form.

7.1 Multi-Modal Grounding

The various lexical acquisition and grounding tasks that were investigated in chapters 5 and 6, and the ARTMAP connectionist architectures that were deployed to tackle these tasks can be considered as somewhat naive if one contemplates the breadth of lexicon that even a young infant is able to understand and use. In both cases the ARTMAP systems could only ground lexical symbols with respect to a single specific sensory domain; colour in chapter 5, and form in chapter 6. Consequently, the nature of the grounded lexicon acquired by each system was extremely constrained. For example the model described in the last chapter was able to acquire some basic meaning of an English term such as 'dog.' However, it would have been quite impossible for this same model to acquire any aspect of meaning *vis-à-vis* other common English terms such as 'red,' 'hot,' and 'fast' simply because it was not endowed with the abilities to sense, or experience those stimuli to which these terms primarily relate, i.e. colour, temperature, and sequences of events or actions. To attempt to teach the meanings of such terms to any artificial model that is without these abilities is akin to the task of teaching a human who lacks hearing the meaning of a term such as 'loud' — a fundamentally difficult, if not impossible problem.

The grounding tasks described in chapters 5 and 6 can be thought of as single mode tasks and consequently the ARTMAP models as single mode systems. These single mode tasks and systems have been the primary focus for the majority of established research within this field[1] and have been important from the perspective of understanding the basic issues that surround automated lexical acquisition and grounding. However, many of these basic issues have now been addressed, and given the results presented in the previous chapters, it is suggested here that this line of research is moving toward maturity. Therefore, it now seems an opportune time to advance beyond these single mode systems and consider more advanced learning architectures that could develop a broader class of meaningful lexicon. To meet this objective new ENLP systems must be endowed with the abilities lacking in those that went before, i.e. they must be able to process input from multiple sensory, and eventually motor modalities. Such systems can be thought of as multi-modal (MM) systems.

[1] The majority of related models and tasks discussed in this thesis (e.g., see chapter 3) fall squarely into the single mode realm.

MM grounding is an obvious step forward but one which has not been addressed adequately in literature, probably due to the relative infancy of this research field. With reference to their single mode grounding models both Dorffner (1992, p.291) and Sales (1996, p.239) make passing reference to architecture scalability, and the ability to relate lexical symbols to more than one sensorimotor modality. However, in neither's work is there any real evidence that any in-depth thought has been given to the impact an upgrade may have in respect to the learning algorithms that each author exercises. Moreover, neither author actually implements and empirically investigates MM grounding.

In chapter 3, two models were described which did empirically investigate lexical acquisition and grounding tasks in the presence of more than a single sensorimotor modality. In the work described by Grumbach (1996) lexical symbols were acquired and grounded in relation to two sensorimotor modalities and in DETE (Nenov & Dyer 1994), this acquisition and grounding occurred in relation to five sensory modalities. The approach taken in each of these systems is now considered in greater detail.

In Grumbach's model (described fully in section 3.2.7), two classes of lexical symbols were acquired and grounded. The first class consisted of symbols such as 'left bend' and 'sharp bend' which were grounded with respect to visual input while the second class consisted of symbols such as 'slow down' and 'turn left,' these being grounded with respect to motor actions. A big problem with Grumbach's model is that it is incapable of autonomous operation – one of the criteria specified in section 3.2.1 – because the author insists on processing linguistic input in two different ways (see, section 3.2.7 for further details). Thus, Grumbach makes decisions on behalf of the model. In a more realistic environment autonomous learning will be required; the designer will not always be around to help the ENLP model out. As far as the model is concerned, all linguistic input is simply delivered via one sensory input stream, and an arbitrary pattern $\mathbf{I}^1$ will be treated no different to a pattern $\mathbf{I}^2$.

The scope of lexical acquisition and grounding demonstrated in Nenov and Dyer's model DETE, (described fully in section 3.2.2) far exceeds that shown in Grumbach's work. Lexical symbols were acquired and grounded with respect to four static sensory modalities, these being colour, form, size, and location. In addition to these static abilities the recurrent nature of the connectionist architecture used in DETE made the acquisition of non-linguistic concepts relating to time dependent occurrences possible. The model was thus able to acquire and ground lexical terms such as ‘bounces’ — a word that relates to motion dynamics. Unlike Grumbach’s model, DETE used a single input channel for all lexical input.

The main weakness of DETE as a MM grounding model is one of developmental inadequacy. In many of the systems sensory modalities no learning functionality is present and the sensory concepts used in grounding are hard coded into the system a priori, e.g. DETE used 3 pre-designed form concepts; circle, square, and triangle. This lack of non-linguistic learning capacity ultimately constrained the scope of lexical terms that could be grounded e.g. It would have been impossible for DETE to obtain a good grounding for a lexical term such as “dog” because this term could only be grounded with respect to one of the models three a priori form concepts[2].

The deficiencies described above have provided the main impetus to seek a better solution to the MM acquisition and grounding assignment. In the following sections a new MM ENLP system based upon ART principles is introduced and investigated.

[2] If a picture of a dog were presented to DETE it would not be recognised as a dog but rather as either a circle, a square, or a triangle.

7.2 An ARTMAP MM Grounding Architecture

The rationale that guides the development of a new MM grounding model is no different to that stated in a previous chapter, i.e. one aims to create a practical system that can be applied in realistic operating environments. As such the 'Model' model criteria specified in chapter 3 remain just as valid when discussing this new task. As such, considering points raised in the preceding section, special attention should be extended to:

- *Autonomy* – In contrast to Grumbach's approach, a single integrated learning architecture is sought. With a single lexical input channel, this type of architecture could be feasibly interfaced to a commercial speech recognition system, which would provide a constant stream of diverse patterns encoded from natural speech.
- *Development* – In contrast to the DETE approach, learning should not be artificially bounded. The ability to construct new internal concepts based on raw sensory data is of utmost importance. It is this capacity to continually categorise sensory experience that actively drives lexical growth.

A further factor that is considered important *vis-à-vis* the properties of an MM model is that of modularity. It is known that specific areas of the human brain are involved with specific sensorimotor activities, e.g. within the field of visual research it is documented that visual input is segregated into colour,form, movement, and depth components, each processed within a specific modular sub-system (Livingstone & Hubel 1988). Moreover, a number of empirical Neurobiological findings indicate language learning occurs through interaction between language modules and other cognitive modules.

It should now be clear from chapters 4 to 6 that ARTMAP networks comply with the aforementioned criteria. Further to this the basic ARTMAP architecture is very modular by design. However, what ARTMAP lacks in its current guise is enough modular components. Given that for a LAG task one ART module must process lexical input, there is only one module left for sensorimotor purposes. Thus, MM grounding is out of the question.

To create a new MM grounding model it seems that the most obvious approach is to extend the standard ARTMAP architecture, enabling the integration of further sensorimotor machinery. Figure 7.1 illustrates a novel architecture that is proposed as the underlying

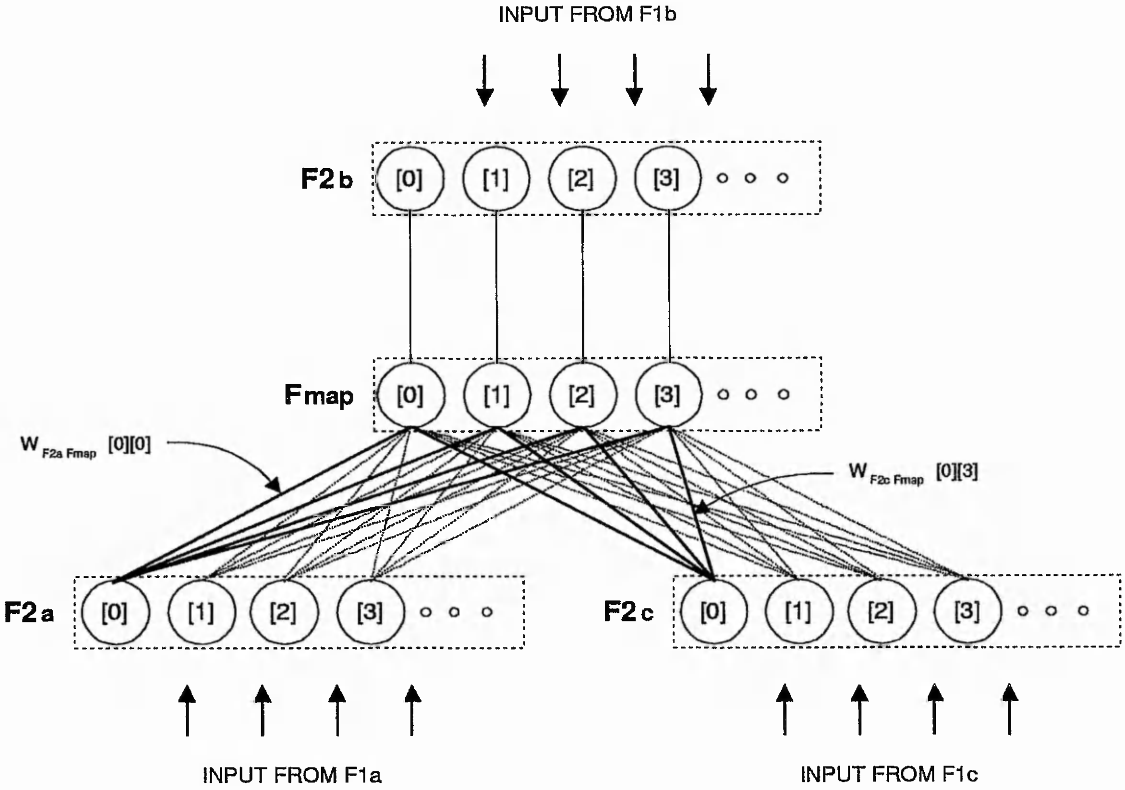

Figure 7.1: Basic Overview of the Multi-Modal Grounding Architecture. The figure shows the category, or F2 layers of each ART module ($F2_a$ - colour, $F2_b$ - linguistic, and $F2_c$ - form) used and graphically indicates how these are connected via a novel three way associative map layer, F_{abc}. The new model is like a fusion of two standard ARTMAP structures. The modules ART$_a$, ART$_b$, and their interconnectivity through the Map Field constituting one such structure and the modules ART$_c$, ART$_b$, and their interconnectivity through the Map Field constituting a second such structure.

“hardware” for a new MM model of lexical acquisition and grounding. The architecture embraces the vision of modularity inherent in the standard ARTMAP architecture and expands upon this through the addition of an extra ART module, ART$_c$. Two learning modules are available for sensorimotor purposes and thus MM grounding can be feasibly realised.

7.2.1 Structure Of The Model

The general structure of the new learning architecture bears a close resemblance to that found in a standard ARTMAP architecture (see chapter 4), however, with the addition of a new module there are specific differences. The full structural details follow :

1. **Input (F1) Layers.** The size of $F1$ is fixed within each ART module. The particular size is dependent on the particular processing requirement e.g. a module processing RGB colour input would have three $F1$ nodes, one stimulated by red activity, one by green activity, and one by blue activity.

2. **Category (F2) Layers.** The size of the $F2$ is dynamic within each module and expands as new categories/concepts are acquired.

3. **F1 ⇔ F2 Connections.** Adaptive pathways lead from each $F1$ node to all $F2$ nodes, and from each $F2$ node to all $F1$ nodes.

4. **Map Field Layer, $\mathbf{F_{abc}}$.** The F_{ab} layer utilised within a standard ARTMAP system to learn a consistent mapping between ART$_a$ and ART$_b$ categories is replaced in the new architecture by the F_{abc} layer. The new F_{abc} layer is required to learn consistent mappings between ART$_a$ and ART$_b$ categories, and between ART$_c$ and ART$_b$ categories.

5. **$\mathbf{F2_a}$ ⇔ $\mathbf{F_{abc}}$ Connections.** In a standard ARTMAP system uni-directional adaptive pathways propagate activity on an $F2$ node in the ART$_a$ module to all nodes in the F_{abc} layer. In line with the modifications discussed in section 4.3.2, these adaptive pathways have been made *bi-directional*, thus facilitating both ‘description,’ and ‘identification’.

6. $\mathbf{F2_b} \Leftrightarrow \mathbf{F_{abc}}$ **Connections.** As in a standard ARTMAP system there is a single bi-directional nonadaptive pathway linking each $F2$ node in the ART$_b$ module to a corresponding node in the F_{abc} layer.

7. $\mathbf{F2_c} \Leftrightarrow \mathbf{F_{abc}}$ **Connections.** The connections between the new ART modules $F2$ layer and the map field are identical to those between the ART$_a$ $F2$ layer and the Map Field; with bi-directional adaptive pathways linking each $F2$ node to all F_{abc} nodes.

8. $\mathbf{F_{abc}} \Rightarrow \mathbf{OS_a}$ **Error Feedback.** As in a standard ARTMAP architecture there is a control mechanism between the Map Field and the ART$_a$ modules orienting subsystem, which facilitates predictive error minimisation.

9. $\mathbf{F_{abc}} \Rightarrow \mathbf{OS_c}$ **Error Feedback.** In the new system there is an additional control mechanism between the Map Field and the ART$_c$ modules orienting subsystem.

7.2.2 Operation Of The Model

Within the new setup the ART$_b$ module is responsible for the processing of all lexical input with the remaining modules ART$_a$, and ART$_c$, dedicated to sensorimotor processing. The connectivity between the Map Field and the three art modules supports associative learning between ART$_a$ and ART$_b$, and also between ART$_c$ and ART$_b$. Therefore, lexical input can be grounded with respect to both sensorimotor domains.

Given the structure of the new model an Adaptive Resonance inspired pseudo cycle of operation can be derived as follows :

```
1. Apply input vectors to ART$_a$, ART$_b$, and ART$_c$.

2. Find the winning $F2$ nodes $J$, $K$, and $L$ in ART$_a$, ART$_b$, and ART$_c$.

3. (a) Compute the associative prediction between $J$ and $K$.

   (b) Compute the associative prediction between $L$ and $K$.

4. (a) i. If $J$ does not predict $K$ then reset ART$_a$ node $J$,
          raise vigilance, find new winning $F2$ node $J'$ in ART$_a$,
          return to step 2 replacing $J$ with $J'$.
```

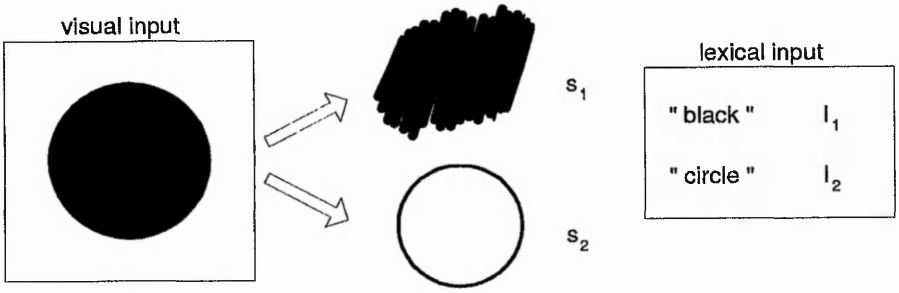

Figure 7.2: An example of typical input available to a MM grounding model. It shows a visual scene containing a single black circle and an accompanying description "black circle".

```
       ii. if $J$ does predict $K$ then update system weights to make
           prediction stronger.
   (b) i. If $L$ does not predict $K$ then reset ART$_c$ node $L$,
          raise vigilance, find new winning $F2$ node $L'$ in ART$_c$,
          return to step 2 replacing $L$ with $L'$.
       ii. if $J$ does predict $K$ then update system weights to make
           prediction stronger.
```

7.2.3 Autonomy and Ambiguous Data

Data ambiguity is an inherent evil in the new MM grounding model due to its architecture. Problems arise because during the course of undertaking an MM grounding task, where various lexical terms must be grounded with respect to either colour or form, a learning system will typically be presented with:

1. a visual scene depicting a coloured object.
2. a linguistic description of the scene.

Consider the example depicted in figure 7.2. In this case the input available to the model is of the form:

$$\{< s_1 >, < s_2 >, < l_1 >, < l_2 >\}$$

with s_1 as an input vector representing the colour black (e.g. the RGB vector - { 0,0,0 }), s_2 as a vector encoding of an n x n binary pixel array, and l_1 and l_2 as vector encodings of the terms "black" and "circle" respectively.

In the new architecture, for the purpose of autonomy, all lexical input is provided through a single lexical channel. The dynamics of the lexical ART module constrain it to processing a single input vector at a time, i.e. each lexical item in ***isolation***. Consequently, processing both the lexical inputs l_1 and l_2 requires two ***sequential*** cycles of learning. During each of these cycles the visual input is constant, therefore, the actual input sets processed by the model are:

$$\begin{array}{ll} \text{cycle 1} & \{< s_1 >, < s_2 >, < l_1 >\} \\ \text{cycle 2} & \{< s_1 >, < s_2 >, < l_2 >\} \end{array}$$

where s_1, and s_2 are respectively supplied as inputs to the learning modules ART$_a$, and ART$_c$ over both learning cycles, l_1 is the input fed to ART$_b$ in the first cycle, and l_2 is the input fed to ART$_b$ in the second cycle.

In chapters 5 and 6 all of the training data provided to the ARTMAP grounding models was of the form

$$\{< s >, < l >\}$$

and was accurate[3], i.e. a given sensory input, s was correctly described by the accompanying lexical input, l. By contrast, in the new model, data can be said to be ambiguous in the sense that it is not known (by the model) whether:

- l_1 is related to s_1.
- l_1 is related to s_2.
- l_1 is related to both s_1 and s_2.

or whether:

- l_2 is related to s_1.
- l_2 is related to s_2.
- l_2 is related to both s_1 and s_2.

[3] a negligible percentage of data used in chapter 5 was actually inaccurate (see section 5.4.1).

7.3 Applying Adaptive Resonance Learning in the Multi-modal Grounding Model

In the single mode ARTMAP grounding systems discussed in chapters 5 and 6 the two major aspects of learning required for lexical acquisition and grounding were :

1. The acquisition of $F2$ concepts based upon generalisations of lexical and sensory stimuli.
2. The creation of associative links between concurrently active sensory and lexical $F2$ concepts.

In the MM system, given the inherent existence of ambiguous training data, the nature these learning requirements are further complicated. Whilst step one remains unchanged, step two takes on an extra dimension :

2. The creation of associative links *only* between active sensory and lexical $F2$ concepts that *consistently reoccur* over a number of cycles of learning.

Thus, given MM input data such as :

Visual Input	Lexical Input
black circle	"black circle"
black square	"this is a black square"
black triangle	"the object is black"

one requires a learning process to capture, through *associative consistency*[4], the relation black $\Leftrightarrow$ "black", and disregard other relations such as black $\not\Leftrightarrow$ "this", and square $\not\Leftrightarrow$ "black".

In section 7.4 a number of adapted control and learning equations especially developed for the MM ARTMAP system are introduced and discussed that facilitate this new learning process. The new equations were developed through an understanding of the various deficiencies of the standard ARTMAP equations in the context of their application to the MM

[4] the visual input black is presented with the term "black" more than with any other term and vice versa.

task. The aim of the current section is to furnish the reader with an insight into these shortcomings and thus provide a foundation for the algorithmic changes that follow.

In order to avoid unnecessary complexity, all descriptions of ARTMAP operations in the ensuing explanations are given in the context of a Gaussian based system. Any specific differences that may emerge from the use of Fuzzy based systems are highlighted.

7.3.1 ARTMAP Learning

In order to describe the deficiencies of the standard ARTMAP equations it will be beneficial to first reiterate the learning process[5]. In a standard bi-modal ARTMAP system the *supervised* learning process can be described as follows :

> Two inputs a^1 and b^1 activate the ART$_a$ and ART$_b$ categories $J = 1$ and $K = 1$. Fast Map Field learning ensures that category $J = 1$ learns to predict *only* category $K = 1$, thus creating an associative link between a^1 and b^1. Next a^2 and b^2 are presented as inputs. Input a^2 activates ART$_a$ category $J = 1$ and b^2 activates a new ART$_b$ category $K = 2$. Following from this state a conflict arises at the Map Field layer. This is because the active ART$_a$ category $J = 1$ predicts the ART$_b$ category $K = 1$ but the currently active ART$_b$ category is not $K = 1$, but $K = 2$. This mismatch prompts the Map Field to send a reset signal to the ART$_a$ orienting sub-system which de-activates the ART$_a$ category $J = 1$. The ART$_a$ search process leads to the activation of a new category $J = 2$. Since $J = 2$ currently makes no prediction there is no mismatch at the Map Field hence Map Field resonance occurs and $J = 2$ learns to predict $K = 2$. Thus, an associative link is created between a^2 and b^2.

The purpose of this supervised learning process can be clarified as follows. Suppose that the ART$_a$ module is processing colour data[6] and that the inputs a^1 and a^2 are examples

[5] A full description of Fuzzy and Gaussian ARTMAP parameters and algorithms is provided in appendix A.

[6] For ease of graphical explanation suppose that this colour input is two dimensional, thus $a^x = (a_1^x, a_2^x)$ and $M = 2$.

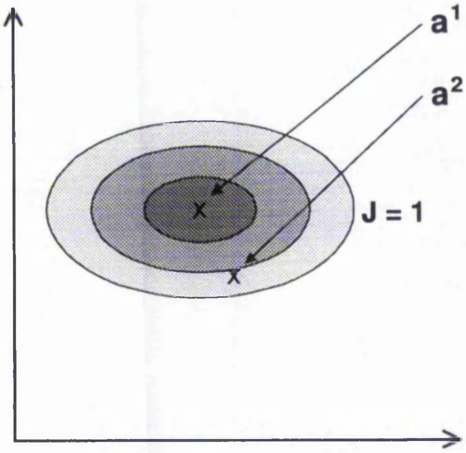

Figure 7.3: The figure shows the ART$_a$ modules two dimensional categorical (weight) space. a^1 and a^2 are shown as points in the input space. The contours are shown to exemplify decreasing levels of match between points in space and the category representaion that $J = 1$ has learnt. Outside of these contours the level of match is such that it is less than ρ_a , and as such mismatch will be registered.

of the colours red and orange respectively. Initially the input a^1 is presented to the ART$_a$ module and since no previous learning has taken place an uncommitted category $J = 1$ is activated and learning ensues. Next, a^2 is presented as input to the system and initially because $J = 1$ is the only committed category it becomes activated, i.e. the network makes an initial hypothesis that a_2 belongs to category $J = 1$. The ART hypothesis testing process now checks if a_2 is sufficiently similar (in vector terms) to the category representation learnt by unit $J = 1$. In this case sufficient similarity is found. Figure 7.3 shows this situation graphically, the Gaussian category $J = 1$ is centered about the vector a^1. The vector a^2 falls within the match region of category $J = 1$.

The current state that has been adopted is wrong. Whilst a^1 is a red input vector and is thus correct in activating category $J = 1$, a^2 is an orange input vector and thus should not activate category $J = 1$ but a new category; an orange item should not belong to the class of items that are red. To overcome this error one applies the *supervised* ARTMAP learning process.

Suppose that the ART$_b$ module processes lexical input and that b^1 and b^2 are two lexical inputs that represent the terms "red" and "orange". Now a^1 is presented to the model in conjunction with b^1. Again a^1 activates category $J = 1$. The lexical input b^1 activates

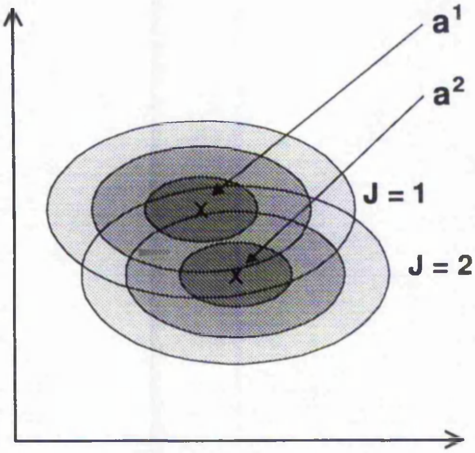

Figure 7.4: The figure shows the match regions of the two Gaussian categories $J = 1$ and $J = 2$ that are a product of the *supervised* ARTMAP learning.

category $K = 1$. Map Field resonance occurs and category $J = 1$ learns to predict $K = 1$, thus forming a link between red (a^1) and "red" (b^1). Next a^2,the orange colour vector, is presented with b^2, the "orange" lexical vector. Since a^2, the orange vector, falls relatively close in sensory space to the vector a^1, it initially activates category $J = 1$. The lexical input b^2 activates category $K = 2$. Category $J = 1$ has learnt to predict $K = 1$, however, the current active ART$_b$ category is $K = 2$. Therefore, there is mismatch at the Map Field layer and a reset signal is sent to the ART$_a$ orienting sub-system. Category $J = 1$ is de-activated, ρ_a is raised to the level of ART$_a$ reset, and a new competitive search process leads to the committal of a new ART$_a$ category $J = 2$. There is no Map Field mismatch between $J = 2$ and $K = 2$, therefore resonance occurs and $J = 2$ learns to predict $K = 2$. Following this, the weights that fan-out from category unit $J = 2$ to the Map Field layer (w_{2k}^{ab}) are set so that in future when a^2 is presented in isolation, it will activate the category $J = 2$ rather than $J = 1$ without the need for *supervision*. Figure 7.4 graphically shows the situation following supervised learning.

7.3.2 Learning With Ambiguous Data

The supervised learning process defined in the previous section is very powerful in that it allows errors to be minimised for any arbitrary set of ART$_a$/ART$_b$ category relations. One main assumption that is made when applying this supervised learning is that the training data is consistent, i.e. that an ART$_a$ input pattern is directly related to the ART$_b$ pattern with which it is concurrently presented. In the MM grounding model the consistent data assumption does not hold. One finds that sensory input is presented to the model in conjunction with lexical input where there is no relation between the two. This ambiguous data leads to a proliferation of pointless sensory categories. The reasons for this are explained below.

Suppose that the two inputs a^1 and a^2 are in fact both (good) examples of the colour red. Because of the nature of the MM grounding task these two inputs are supplied in conjunction with the two ART$_b$ inputs b^1 and b^2 respectively, where b^1 represents the term "red" and b^2 represents "circle" (a term that is totally unrelated to the domain of colour). The input a^1 activates category $J = 1$ and b^1 activates category $K = 1$. With fast Map Field learning ($\beta = 1.0$) $J = 1$ learns to predict *only* $K = 1$ and vice versa. Thus, an associative link is created between red (a^1) and "red" (b^1). Next the input a^2 activates category $J = 1$, which is correct since a^1 is an example of the concept red. However, b^2 is unlike b^1 and hence activates a new category $K = 2$. This causes Map Field mismatch and thus category $J = 1$ is de-activated. A new ART$_a$ category $J = 2$ is selected to code a^2.

Suppose two further inputs a^3 and a^4, again both (good) examples of the colour red, are presented to the system in conjunction with the two inputs b^1 i.e. "red" and b^3, which represents the term "square". The input a^3 activates category $J = 1$ and in the ART$_b$ module b^1 activates category $K = 1$. Since $J = 1$ predicts $K = 1$ Map Field resonance occurs. The weight vector μ_1^a is updated to take into account the input a^3. The input a^4 activates category $J = 1$, which is correct since a^4 is an example of the concept red. However, b^3 activates category $K = 3$. This causes Map Field mismatch and thus category $J = 1$ is de-activated. A new ART$_a$ category $J = 3$ is selected to code a^4.

The inputs a^1 through a^4 should all have been coded on the same ART$_a$ category. However, because of the Map Field dynamics three ART$_a$ categories have been used to code these inputs and only one of these has encoded a satisfactory associative connection to the ART$_b$ network.

7.3.3 Fast vs. Slow Map Field Learning

In the previous two sections the ARTMAP calculations were described from the perspective of *fast* Map Field learning (the parameter $\beta_{map} = 1.0$). With fast Map Field learning an active ART$_a$ category, e.g. $J = 1$, learns to predict *only* the current active ART$_b$ category, e.g. $K = 1$. The associative links between $J = 1$ and all other ART$_b$ categories (where $K \neq 1$) are set to zero. Following this learning, if an input a^x activates $J = 1$ and a concurrently presented ART$_b$ input b^x activates a category K where $K \neq 1$ a conflict occurs and mapfield mismatch is registered, as in the above case. Algorithmically this is because:

$$w^{ab}_{JK}\ (= 0.0\,) \quad < \quad \rho_{map}$$

With slow(er) Map Field learning ($\beta_{map} < 1.0$) an active category $J = 1$ again learns to maximally predict ART$_b$ category $K = 1$. However, the associative links between $J = 1$ and the other ART$_b$ categories are not nullified but simply reduced in strength, e.g. $w^{ab}_{Jk} = 0.8$, for all k's where $k \neq K$. Following a single presentation of slow learning if an input a^x activates $J = 1$ and a concurrently presented ART$_b$ input b^x activates a category K where $K \neq 1$ *no* conflict occurs and Map Field resonance is registered. This is because now:

$$w^{ab}_{JK}\ (= 0.8\,) \quad \geq \quad \rho_{map}$$

Therefore, the application of slow Map Field learning has averted category proliferation in the presence of data ambiguity. If slow Map Field learning is applied to the example data defined in the previous section (7.3.2), then, when a^2 is presented with b^2 ("circle") there is no longer mismatch between $J = 1$ and $K = 2$. Therefore, input a^2 is coded by category $J = 1$ as is required. In the same way a^4 is no longer coded on a new category $J = 3$, it now wins as required on $J = 1$.

Following from the above it would seem that the properties of a slow learning system are well suited to the MM task domain. Unlike in a fast learn system where any relation inherent within the training data is learnt by the system within one presentation, in a slow learn system, it appears that only those relations that occur consistently throughout the whole of the training data are acquired. Although this observation holds true, it does so only to a certain point. One can explain this position fully through an extension of the learning process described in the previous example.

Suppose that the data in the previous example is presented again to the network a number of times, e.g. a second, third, fourth, ... epoch. In each epoch a^1, a^2, a^3, and a^4 activate ART$_a$ category $J = 1$ and in each epoch this category is active concurrently with the ART$_b$ categories $K = 1$, $K = 2$, $K = 1$, and $K = 3$. Since $J = 1$ is continually being presented in conjunction with $K = 2$ and $K = 3$ less number of times than it is with $K = 1$ the weight values w^{ab}_{12} and w^{ab}_{13} begin to diminish considerably and to a greater degree than the w^{ab}_{11} weight value, thus :

$$w^{ab}_{12} = w^{ab}_{13} < w^{ab}_{11}$$

Although this system state embodies exactly the relations that one hoped would be learnt, i.e. that an input (a^x) which activates $J = 1$, is more likely related to the term "red" and less to the terms "circle" or "square", there are problems in terms of further training. The weight values w^{ab}_{12} and w^{ab}_{13} eventually get so low that when an a vector that activates $J = 1$ is presented to the system with the term "circle" or "square" the ARTMAP mismatch test registers predictive conflict, i.e.

$$w^{ab}_{JK}\ (= small) \quad < \quad \rho_{map}$$

Following this result extra ART$_a$ categories are recruited by the system, the problem of category proliferation returns.

One obvious solution to this particular problem is to make learning very slow ($\beta_{map} \approx 0.0$) and make the Map Field vigilance level very small ($\rho_{map} \approx 0.0$) thereby effectively disabling the ARTMAP Match Tracking process. Although this action will stop the process of MM category proliferation, e.g. the creation of many overlapping red concepts, it will also completely block the creation of any other valid concepts (as described in section 7.3.1), e.g. the creation of a blue concept. As discussed below, this failing will consequently cause greater problems.

Training Cycle	ART$_a$ Input	ART$_b$ Input	ART$_c$ Input
1	a^1 (red)	b^1 ("red")	c^1 (circle)
2	a^1 (red)	b^2 ("circle")	c^1 (circle)
3	a^2 (blue)	b^3 ("blue")	c^1 (circle)
4	a^2 (blue)	b^2 ("circle")	c^1 (circle)

Table 7.1: Training data used in a simple experiment (see main text for details).

Suppose that in a different training example a^1 and a^2 are examples of red and blue respectively. Given that both of these inputs are supplied to the learning system in conjunction with a single ART$_c$ input, which represents an example of a circle, the training data available to the ART$_a$ and ART$_b$ can be derived as shown in table 7.1.

In a slow learn system a^1 and a^2 will continually activate category $J = 1$ (since Match Tracking can't create a new category for a^2). Referring to the data in table 7.1 one can see that $J = 1$ will be active over four cycles in conjunction with the categories $K = 1$, $K = 2$, $K = 3$, and $K = 2$. In this example since $J = 1$ is continually being presented in conjunction with $K = 1$ and $K = 3$ less number of times than it is with $K = 2$ the weight values w^{ab}_{11} and w^{ab}_{13} diminish to a greater degree than the w^{ab}_{12} weight value, thus :

$$w^{ab}_{11} = w^{ab}_{13} < w^{ab}_{12}$$

One can see that the system has now learnt the entirely wrong relation, i.e. that an input (a^x) which activates $J = 1$ is ultimately related to the term "circle". There is also one further problem. Although the vectors a^1 and a^2 are actually activating the same category $J = 1$ it is possible that they may be situated some distance apart in the sensory vector space. Following ART learning category $J = 1$ will acquire a prototype vector that is based upon some generalization of these two vectors, thus the centre of the category will reside somewhere centrally between these vectors . Figure 7.5 shows this situation graphically.

The problem with this is that the region of space that resides between these vectors, and where the centre of the category is situated, may well have little to do with either red or blue and more to do with another sensory category, e.g. purple. Thus, one has created a relation between purple and "circle", which is of little use.

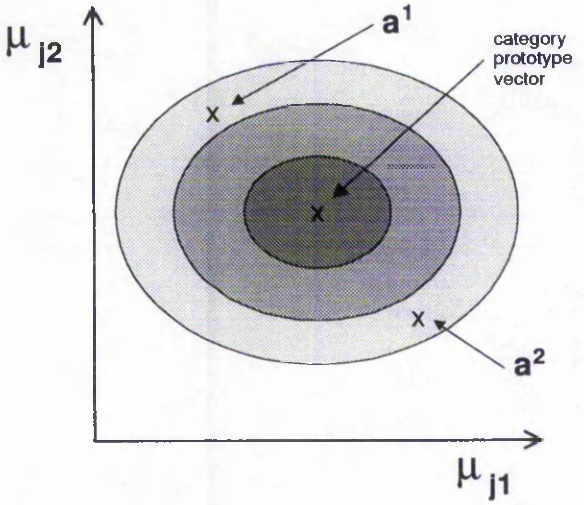

Figure 7.5: The placement (centre) of the Gaussian category in space is based upon some generalization of the points a^1 and a^2.

The inability to create valid concepts such as blue is a serious failing since the issue of concept development is vital in a grounding system. This point cannot be stressed enough. If an artificial system cannot form the correct concept boundaries within sensory space then it has no way of securing any 'meaningful' meaning[7] for the lexical terms it acquires.

The discussions above have considered how both fast and slow learn standard ARTMAP systems can be applied to an MM grounding task where training data is inherently ambiguous. It has been shown that, given this data, the learning task is non-trivial and that the standard ARTMAP approach is not sufficient in either its fast or slow learn guise.

[7] By virtue of some reference to a region in sensory space one might argue that an acquired lexical term will have some meaning. However, since this conceptual region does not directly correspond to the region that a human observer may expect, the 'understanding' the artificial system has for this term will differ from that the human observer would typically have. Consequently to the human observer the meaning that the artificial system has acquired is meaningless.

7.4 A Novel Adaptive Resonance Algorithm for Multi-Modal Grounding

In light of the failings described above, a fundamental and important contribution of the work described in this thesis has been to develop an adapted method of learning that can be applied in a MM context. In the following sections the details of a number of adaptations to the standard ARTMAP learning algorithms are presented and described.

7.4.1 Addressing Map Field Controller Problems

One of the main problems that emerged from the examinations of standard learning set out in the previous sections centered around the process of Match Tracking. To reiterate, the problem is as follows. On one hand the process is vital, without it there is no way to resolve possible predictive errors that are highlighted from the presentation of *accurate* training data, but on the other hand the process fails to deliver when confronted with ambiguous data, as good associative mappings are corrupted.

The onset of Match Tracking is governed in an ARTMAP system by the Map Field Controller. Consequently, in order to redress this issue modifications have been made to the Map Field control equations. The new equations work toward[8] overcoming the problem of ambiguous data by taking into account the current activity in all three ART modules. The new equations are introduced below and then their application in in the context of a MM training task is considered.

[8] The full solution also additionally depends on (1) an adapted Map Field learning algorithm, and (2) some new initialisation steps. These variations are expanded upon in later sections of the chapter.

RULE 1

if $w^{ab}_{JK} \geq \rho_{map}$ & $w^{cb}_{LK} \geq \rho_{map}$

- register ART$_a$ ⇔ ART$_b$ Match and ART$_c$ ⇔ ART$_b$ Match.
- initiate ART$_a$, ART$_b$, and ART$_c$ Resonance and Learning.
- initiate ART$_a$ ⇔ ART$_b$ and ART$_c$ ⇔ ART$_b$ Map Field Resonance and Learning.

RULE 2

else if $w^{ab}_{JK} \geq \rho_{map}$ & $w^{cb}_{LK} < \rho_{map}$

- register ART$_a$ ⇔ ART$_b$ Match and ART$_c$ ⇔ ART$_b$ Mismatch.
- initiate ART$_a$, and ART$_b$ Resonance and Learning.
- initiate ART$_a$ ⇔ ART$_b$ Map Field Resonance and Learning.

RULE 3

else if $w^{ab}_{JK} < \rho_{map}$ & $w^{cb}_{LK} \geq \rho_{map}$

- register ART$_a$ ⇔ ART$_b$ Mismatch and ART$_c$ ⇔ ART$_b$ Match.
- initiate ART$_b$, and ART$_c$ Resonance and Learning.
- initiate ART$_c$ ⇔ ART$_b$ Map Field Resonance and Learning.

RULE 4

else

- register ART$_a$ ⇔ ART$_b$ Mismatch and ART$_c$ ⇔ ART$_b$ Mismatch.
- initiate ART$_a$, and ART$_c$ Match Tracking.

Training Cycle	ART$_a$ Input	ART$_b$ Input	ART$_c$ Input
1	a^1 (red)	b^1 ("red")	c^1 (circle)
2	a^1 (red)	b^2 ("circle")	c^1 (circle)
3	a^1 (red)	b^1 ("red")	c^2 (square)
4	a^1 (red)	b^3 ("square")	c^2 (square)

Table 7.2: Training data used in another experiment (see main text for details).

To avoid a proliferation of unnessesary committed categories in a Gaussian based MM model one also has to make slight amendments to the Match Tracking process. As per standard Match Tracking, vigilance in both the ART$_a$ and ART$_c$ modules is raised to a level whereby resonance of J and L is inhibited and further ART processing either results in:

1. Resonance of a different committed category, e.g. J' or L'
2. Resonance of a new uncommitted category if no committed categories resonante, i.e. pass the Matching test.

If the case arises where Match Tracking would result in resonance of a different committed unit in one module and resonance of a new uncommitted category unit the other module then vigilance in the latter module is lowered to its previous level so that resonance of the original unit is reinstated. Thus one ends up in a situation where J' resonates simultaneously with L, or J with L'.

The application of these equations is now examined by reconsidering the simple training situation introduced in section 7.3.2. A high level description of the data that is applied to the MM model over four cycles of training is provided in Table 7.2.

The initial cycles of activation, prediction, and learning that emerge from applying the adapted algorithm follow a similar pattern to those already described in section 7.3.3. To re-iterate, in the first learning cycle the inputs a^1, b^1, and c^1 activate the categories $J = 1$, $K = 1$, and $L = 1$ respectively. Since both the weights w^{ab}_{JK} and w^{cb}_{LK} are maximal, Map Field control initiates resonance and learning. Following learning the Map Field weights w^{ab}_{Jk} and w^{cb}_{Lk}, for all k's where $k \neq 1$, are reduced in strength by small proportion. In the next cycle the inputs activate $J = 1$, $K = 2$, and $L = 1$. Resonance and learning prevail

	J=1 K=1 L=1	J=1 K=2 L=1	J=1 K=1 L=2	J=1 K=3 L=2	Total weight change
w_{11}^{ab}		-		-	-2
w_{12}^{ab}	-		-	-	-3
w_{13}^{ab}	-	-	-		-3
w_{11}^{cb}		-			-1
w_{12}^{cb}	-				-1
w_{13}^{cb}	-	-			-2
w_{21}^{cb}				-	-1
w_{22}^{cb}			-	-	-2
w_{23}^{cb}			-		-1

Table 7.3: Table shows how Map Field weights change over four learning cycles. A (-) shows that a weight has decreased in value following the learning in any given cycle. The final column of the table shows the total change for each weight.

and in this step the Map Field weights w_{Jk}^{ab} and w_{Lk}^{cb}, for all k's where $k \neq 2$, are reduced in strength. In the third cycle $J = 1$, $K = 1$, and $L = 2$ are active. Again resonance and learning are initiated. In this cycle the Map Field weights w_{Jk}^{ab} and w_{Lk}^{cb}, for all k's where $k \neq 1$, are reduced in strength. Finally, in the forth cycle $J = 1$, $K = 3$, and $L = 2$ are active. Following resonance and learning the Map Field weights w_{Jk}^{ab} and w_{Lk}^{cb}, for all k's where $k \neq 3$, are reduced in strength. Table 7.3 shows graphically how the various Map Field weights change over these four cycles of learning.

In order to demonstrate the beneficial properties of the new Map Field algorithm an assumption is now made as follows. ρ_{map} is set such that its value is greater than any Map Field weight which has had its value decreased more than three times. Following this assumption a second epoch of learning is now considered.

In the first cycle of the second epoch the inputs again activate $J = 1$, $K = 1$, and $L = 1$. Since w_{11}^{ab} and w_{11}^{cb} are both still greater than ρ_{map}, Match is registered and learning ensues. The Map Field weights w_{Jk}^{ab} and w_{Lk}^{cb}, for all k's where $k \neq 1$, are reduced in strength. Consequently, the weights w_{12}^{ab} and w_{13}^{ab} have now decreased in value *four* times. In the second cycle the inputs a^1, b^2, and c^1 activate the categories $J = 1$, $K = 2$, and $L = 1$. Since w_{JK}^{ab} is now less than ρ_{map} an ART$_a$⇔ART$_b$, Mismatch is registered. At this point

	J=1 K=1 L=1	J=1 K=2 L=1	J=1 K=1 L=2	J=1 K=3 L=2	J=1 K=1 L=1	J=1 K=2 L=1	J=1 K=1 L=2	J=1 K=3 L=2	Total weight change
w_{11}^{ab}		-		-					-2
w_{12}^{ab}	-		-	-	-		-		-5
w_{13}^{ab}	-	-	-		-		-		-5
w_{11}^{cb}		-				-			-2
w_{12}^{cb}	-				-				-2
w_{13}^{cb}	-	-			-	-			-4
w_{21}^{cb}				-				-	-2
w_{22}^{cb}			-	-			-	-	-4
w_{23}^{cb}			-				-		-2

Table 7.4: After two epochs of learning the state of Map Field weights is as above. It can be seen from the table that in learning cycles 2 and 4 (epoch 1) the value of w_{11}^{ab} was reduced. However, in cycles 6 and 8 (epoch 2) this weight reduction is avoided and as such the total weight reduction after two epochs remains equal to that registered after 1 learning epoch (see table 7.3).

the application of the standard Map Field control algorithm would initiate the process of Match Tracking and a new ART$_a$ category would become committed. By applying the new algorithm, this Match Tracking scenario can be avoided. Mismatch is still registered but because w_{LK}^{cb} is greater than ρ_{map} Match Tracking is not initiated. In addition the new algorithm prevents ambiguous data from damaging other knowledge within the system. This prevention can be seen in table 7.4 where Match Tracking can be seen to block w_{11}^{ab} weight reduction. This is important since this weight holds the association between the ART$_a$ category that encodes the colour red and the ART$_b$ category that encodes the term "red", and to effectively represent this relation must be kept at a value greater than ρ_{map}.

Two important factors that were implicit in the example described above were :

1. **Slow Map Field Learning** – It was noted above that when weights were reduced in value they were so by a small proportion. The new Map Field control algorithm will only function appropriately if it is implemented in conjunction with slow Map Field learning. In a fast learn system the weights between an active sensory category and all of the non active lexical categories e.g. w_{Jk}^{ab}, where $k \neq K$ have their value reduced to below the Map Field vigilance threshold in a single learning cycle. The

adapted algorithm is dependent on the associative consistency within the input data that allows some weights to remain above the ρ_{map} threshold whilst others fall below it.

2. **Categorical Separation** - It was assumed in the example above that ARTc input patterns c_1 and c_2 would naturally activate different category units $L = 1$ and $L = 2$. In the initial stages of learning with certain network parameters this assumption may not hold. Solutions to this are addressed in section 7.4.3 below.

7.4.2 Addressing Map Field Learning

In the last section the importance of slow Map Field learning was addressed. In this section a number of adaptations to the standard slow learn algorithm are introduced that complement the new Map Field control system. These changes centre around the addition of three new parameters η, ϕ and β', where:

1. η and ϕ facilitate *predictive reinforcement*
2. β' facilitates *confident fast learning.*

An explanation of these abilities follows the specification of the new learning mechanism.

Adapted Learning Between ART$_a$ and ART$_b$

$$w_{Jk}^{ab} = (1-\beta')w_{Jk}^{ab} + \beta'((Y_k^b \wedge w_{Jk}^{ab}) + \eta)$$

where

$$\beta' = \begin{cases} 1.0 & \text{if } w_{JK}^{ab} = 1.0 \text{ and } w_{LK}^{cb} \leq \rho_{map} \\ \beta & \text{otherwise} \end{cases} \tag{7.1}$$

and

$$\eta = \begin{cases} 0.0 & \text{if } \phi < w_{JK}^{ab} \geq 1.0 \\ \frac{Y_k^b \wedge w_{Jk}^{ab}}{\beta'} & \text{otherwise} \end{cases} \tag{7.2}$$

Adapted Learning Between ART$_c$ and ART$_b$

$$w^{cb}_{Lk} = (1-\beta')w^{cb}_{Lk} + \beta'((Y^b_k \wedge w^{cb}_{Lk}) + \eta)$$

where

$$\beta' = \begin{cases} 1.0 & \text{if } w^{cb}_{LK} = 1.0 \text{ and } w^{ab}_{JK} \leq \rho_{map} \\ \beta & \text{otherwise} \end{cases} \tag{7.3}$$

and

$$\eta = \begin{cases} 0.0 & \text{if } \phi < w^{cb}_{LK} \geq 1.0 \\ \frac{Y^b_k \wedge w^{cb}_{Lk}}{\beta'} & \text{otherwise} \end{cases} \tag{7.4}$$

where β is the standard ART learning rate parameter (see appendix A) and Y^b is the vector of activation across the ART$_b$ F_2 layer. ϕ is described below.

Predictive Reinforcement

In the standard approach to learning all Map Field weight values are initialised with unitary value and then only allowed to decrease as learning progresses. If categories J and K are active then following the standard leaning rule (see appendix A, equation A.19), the weight between the active categories w^{ab}_{JK} remains unchanged whilst the weights between the active ART$_a$ category J and all of the non-active ART$_b$ categories, i.e. w^{ab}_{Jk} for all k's where $k \neq K$, are reduced in value by multiplying each weight's current value with the quantity $(1-\beta)$.

In the new approach the weight reduction process remains unchanged, however, the value for the weight between the active categories J and K is now allowed to increase by the quantity η. This allows the model to maintain strong associations between units that are regularly active together and also provides the basis for confident fast learning, which is described in the following section.

As can be seen in equations 7.2 and 7.4 the value η is dependent on the current value of the of the weight that connects the active ART$_a$ or ART$_c$ category with the active ART$_b$

	J=1 K=1 L=1	J=1 K=2 L=1	J=1 K=1 L=2	J=1 K=3 L=2	Total weight change
w_{11}^{ab}	+	-	+	-	0
w_{12}^{ab}	-	+	-	-	-2
w_{13}^{ab}	-	-	-	+	-2
w_{11}^{cb}	+	-			0
w_{12}^{cb}	-	+			0
w_{13}^{cb}	-	-			-2
w_{21}^{cb}			+	-	0
w_{22}^{cb}			-	-	-2
w_{23}^{cb}			-	+	0

Table 7.5: Shows how predictive reinforcement affects the change in Map Field weights over four learning cycles. A (-) shows that a weight has decreased in value, a (+) that a weight has increased in value, following the learning in any given cycle.

category. If this weight currently has a maximal unitary value then η is set to a value of zero, i.e. the weights value will not increase. This is important since all Map Field weights must be kept within certain bounds:

$$w_{*b}^{*k} \in [0, 1]$$

The value of η is also set to zero when the weights value reduces below a threshold ϕ, where $\phi \in [0, 1]$. Therefore, once a weight between two Map Field categories falls below this criteria it can no longer rise in value.

Outside of the above constraints η is always set to a value of $(Y_k^b \wedge w_{*k}^{*b})(\beta'^{\,-1})$. This value has been chosen in order that the weight in question will have its value increased by the proportion with which it last reduced. To the best of the authors knowledge such a weight increase has never before been implemented in conjunction with an Adaptive Resonance model. Table 7.5 shows a reworking of the weight changes described in section 7.4.1 (summarised in table 7.3) with predictive reinforcement in place.

Confident Fast Learning

Generally for the new model to function as desired quite slow learning is a requirement. However, in certain *confident* situations it is possible to override this default operation and increase the speed of learning.

Suppose the target of learning is for a certain ART$_a$ category J to predict an ART$_b$ category $K = 1$. During a training period J is presented in conjunction with $K = 1$ slightly more than it is with any other k. Therefore, with predictive reinforcement w^{ab}_{JK} retains its initial unitary value whilst the w^{ab}_{Jk} weights diminish slightly. Also during training ART$_c$ category L is been presented with $K = 2$ significantly more than it is with $K = 1$ and consequently w^{cb}_{LK}, where $K = 1$ has diminished considerably to below ρ_{map}, i.e. there is no relation between L and $K = 1$.

Following the above, three inputs are presented to the model and these activate the categories J, $K = 1$, and L. Since w^{ab}_{JK} is maximal and w^{cb}_{LK} is minimal one can infer that $K = 1$ is unambiguously related to the the ART$_a$ domain. Consequently, the assumption must be that J *is* related to $K = 1$ and thus whilst w^{ab}_{JK} remains maximal the weights w^{ab}_{Jk} can be reduced to zero. Equations 7.1 and 7.3 enable this functionality, when $\beta' = \beta$ slow learning holds but when $\beta' = 1.0$ fast learning intervenes.

Backward Compatibility

An important aspect of the new Map Field learning algorithm is its backward compatibility in relation to the standard ARTMAP model. In the MM application that is discussed throughout this chapter, i.e. that task of concurrently grounding terms relating to either colour or form, it is doubtful that activity could be present in one sensory module (colour), without there being activity in the other module (form), since both of these stimuli would typically registered through the same sensory substrate. However, in alternative target domains it is quite possible that the various sensory stimuli will not always be coupled in such a way. Consequently, the problem of ambiguous data vanishes and the standard learning approach will suffice. In such a situation, the new Map Field learning algorithm will function with equivalence to the standard algorithm.

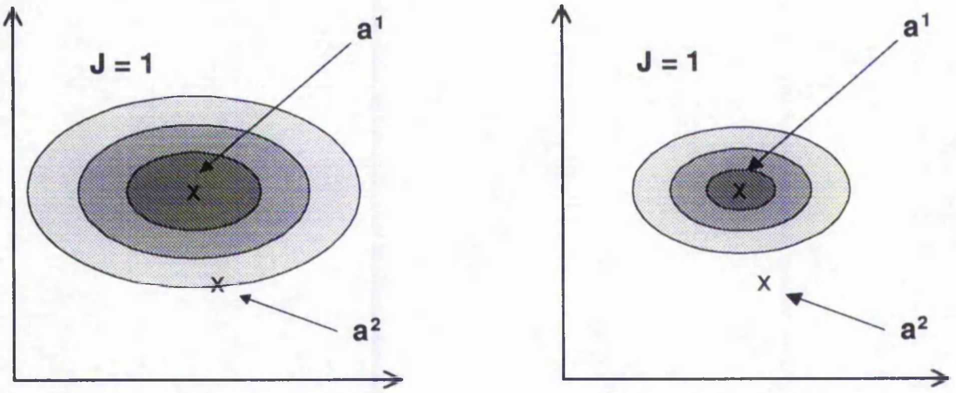

Figure 7.6: Figures (a) and (b) show the two dimensional categorical (weight) spaces for two separate unsupervised ART systems. In both cases input vectors a^1 and a^2 are shown as points projected onto this space. In (a) a low level of ART vigilance effects a large match region for the category $J = 1$. As a result input a_2 is also seen as belonging to this category. In (b) a higher level of vigilance effects in a much smaller match region. In this case a_2 mismatches $J = 1$ which leads to the choice or creation of a new category $J = 2$.

7.4.3 Categorical Separation

One final issue, mentioned briefly above, that one must address is that of initial categorical separation. Earlier in the chapter (see section 7.3.1 and figure 7.3) it was shown that ARTMAP supervised learning, effected by Map Field processing, can overcome problems where two distinct inputs can be categorised as the same thing; the result being that two categories are correctly created rather than just one. Such operation was, however, *falsely* assumed in the preceding examples where c^1 (circle) and c^2 (square) were said to respectively activate units $L = 1$ and $L = 2$ in the initial learning trials. This is because such operation in the initial cycles of learning is dependent on either:

- A maximal Map Field learning rate.
- Maximal ARTMAP vigilance.
- A high learning rate coupled with a high ARTMAP vigilance

and such parameterisations are simply not appropriate when faced with ambiguous data.

One way to overcome this issue is by setting higher levels of vigilance in each of the ART modules. As shown in figure 7.6 this results in a higher likelihood that two distinct inputs will be independently categorised. The main problem with this approach is that it can lead to a 'greedy' system that requires many extra ART processing units and does not comply with the ART philosophy of maximising performance whilst minimising system resources.

Another method, and that which is preferred here, works by initialising a small number of categories, which are marked as committed, prior to any learning. These are initialised with randomised weight vectors, thus effectively distributing them within the weight space. In a Fuzzy-based network this practice may lead to a larger than normal population of categories since Fuzzy dynamics minimise LTM weight values. This is not a concern for Gaussian-based networks since the mean weight values can go up as well as down.

7.5 Evaluating the Adapted ARTMAP Algorithm

In order to assess the new ARTMAP algorithms a simple MM task has been devised. Following from the work in the previous two chapters this task considers lexical acquisition and grounding in the presence of two forms of sensory input and one lexical input:

- RGB colour patterns.
- 144-bit 12x12 pixel visual patterns.
- 10 bit lexical patterns

To create a training data set five of Sales' pixel based images were selected (dog, cat, dolly, clock, & chair (see figure 6.1 in chapter 6)) and five colours (orange, red, yellow, blue, green). Fifteen variations of each of the five colours were employed thus generating a training set containing 75 differently coloured objects. The images were matched to colours in a uniform way, i.e. there were 3 yellow chairs, 3 yellow dogs etc.

A single MM model was employed that was made up of three Gaussian ART modules — ART$_a$ processing colour data, ART$_b$ lexical data, and ART$_c$ form data. In each of the modules, 10 ART category (F_2) units were initialised with random weight configurations. The main ART parameters[9] were set as follows:

[9] For a full description of ART parameters refer to Appendix A.

$rho_m ap$	=
$rho_b ase$	=
β	=

Table 7.6: Parameters used in Investigations

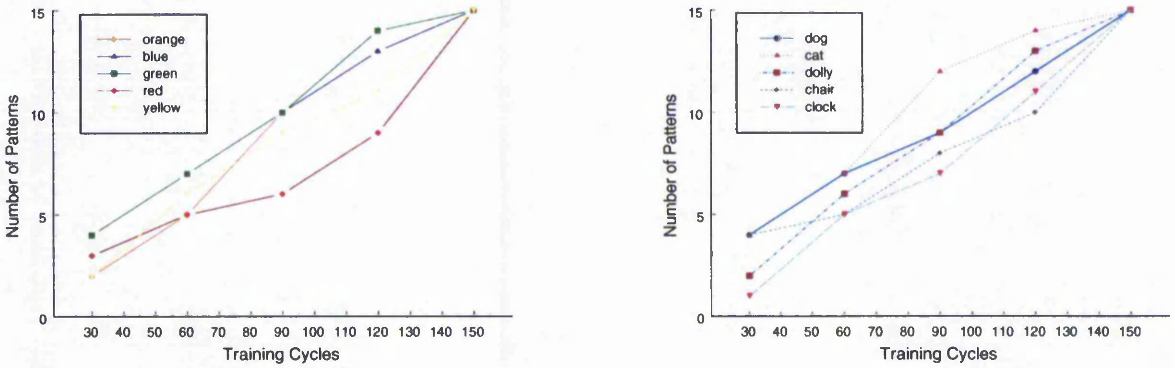

Figure 7.7: The two graphs show the frequency with which data from the training set was presented to the MM grounding model. The left- and right-hand graphs show how many items of a particular colour and how many of a particular form were presented in various training training periods.

Two epochs of training were undertaken, during which each coloured object was presented in turn with each of the words that described it, e.g. Yellow Dog. Thus, in each epoch there were a total of 150 input cycles (75 coloured objects described by 2 lexical entities). The training data was presented in a random order, the frequency of which can be seen in the graphs shown in figure 7.7. The graphs show that over a full epoch of learning 15 examples of each colour and each object were provided to the learning system.

Upon presentation of an input triplet normal Gaussian ART processing within each of the 3 modules ensued. Following this the winning (resonating) units J, K and L and in the three ART modules was noted. During the course of training a total of five units in each of non-lexical modules were found to resonate – units 0, 2, 5, 8, and 9 in ART$_a$ and units 4, 5, 6, 7, and 8 in ART$_c$. Figure 7.8 shows unit resonance activity for the full 150 cycles of processing in the first epoch.

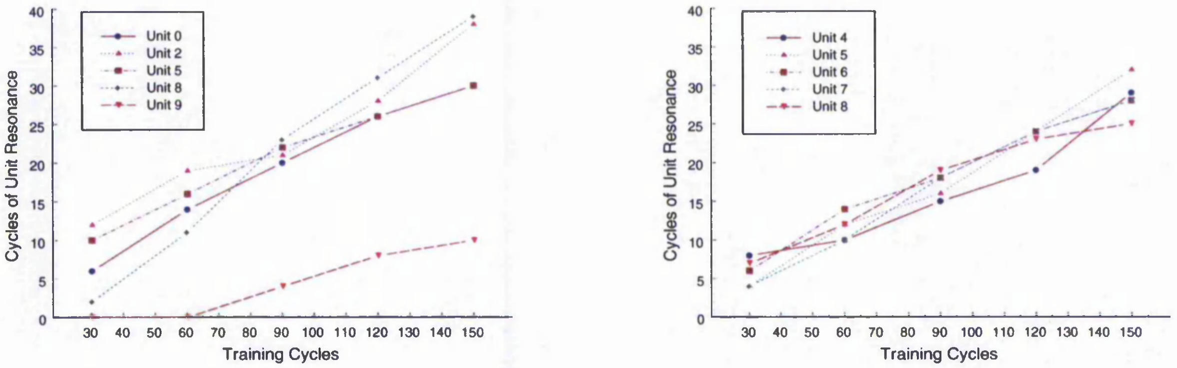

Figure 7.8: ART unit resonance data. The left side graph provides a record of F_2 unit resonance in the colour processing module, ART_a. The right side graphs provides similar data in respect of the form processing module, ART_c.

Following unit resonance the new Map Field rules were applied and learning was undertaken. Following this, the Map Field weights connecting winning (resonating) units in the ART_a and ART_c modules to those in the ARTb module were examined to note what changes have occurred. The graphs presented below in figures 7.9 through 7.18 provide the detail of all of these changes.

Figure 7.9 shows how the weights that connect F_2 unit 0 in ART_a to all F_2 units in ARTb vary over the course of learning. In the first 7 cycles one can see zig-zag patterns that show certain weights decreasing and increasing in value over consecutive cycles. In cycle 1 the value of the weight which effectively binds ART_a unit 0 to the lexical term 'blue' stays maximal whilst all others decrease. This is followed in cycle 2 by an increase in the weight value which binds the unit to the term 'dog' and a decrease in all the other weight values. Cycles 3, 4, and 5 see the weight values binding the unit to 'blue','dolly', and 'blue' respectively increase whilst all others fall.

The zig-zap phenomena is consistent with what one would expect in the early cycles of learning as the MM model has not yet seen enough consistency between resonance in the various ART modules and thus cannot yet dismiss many potential associations. In technical terms, a number of weight values have yet to fall below the level of ρ_{map}, therefore, the new Map Field rules 2, 3, and 4 can not yet be applied to aid learning and overcome the data

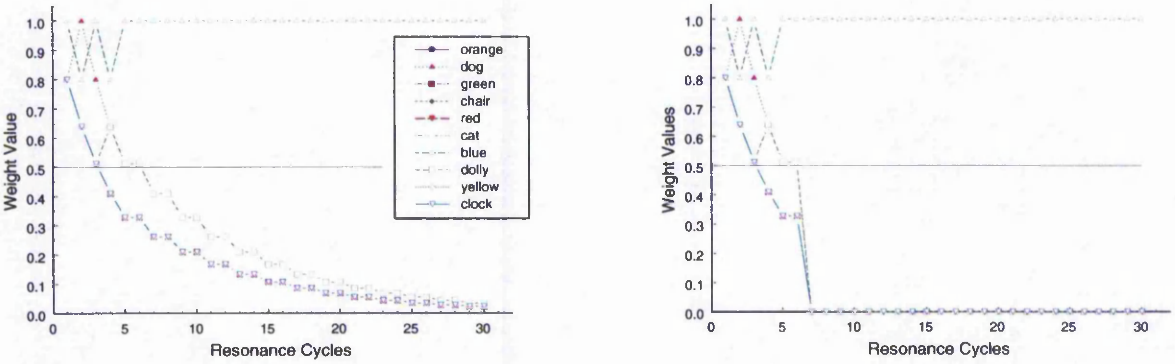

Figure 7.9: The two graphs show how the Map Field weights that connect F_2 unit 0 in ART$_a$ to all F_2 units in ARTb vary over the course of learning. The key provides detail of ART$_b$ units in terms of the lexical term they encode. Hence, each coloured graph line shows the association between unit 0 and a lexical term. The black horizontal line provides a reference of the value of ρ_{map} ($= 0.5$). The left-hand graph shows how the weights change when β' is always equal to β. The right-hand graph shows how weights change when allowing *beta'* to equal 1 under certain conditions, i.e. adding a confident fast learn component to the mix.

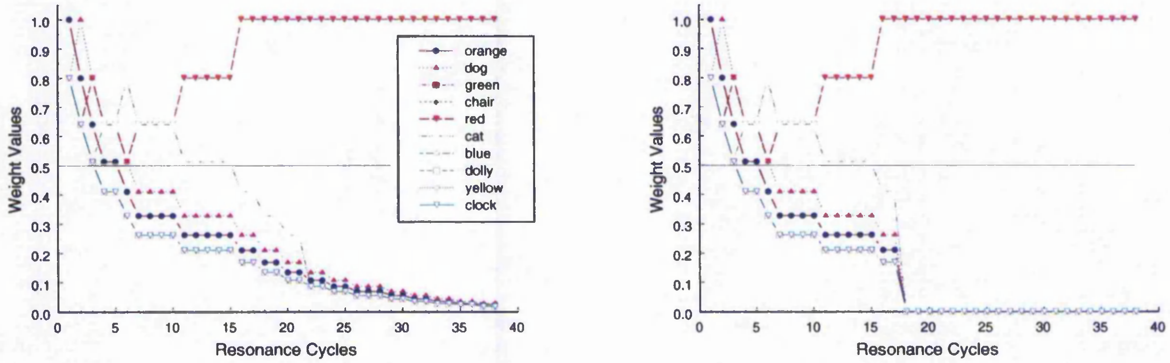

Figure 7.10: The two graphs show how the Map Field weights that connect F_2 unit 2 in ART$_a$ to all F_2 units in ARTb vary over the course of learning.

ambiguity problem. Hence, in cycles 2 and 4 the weight values that connect ART$_a$ unit 0, a colour unit, with the terms 'dog' and 'dolly' are unfortunately reinforced.

Cycle 6 of ART$_a$ unit resonance sees a different pattern emerge as all of the weights keep a constant value. This type of pattern is consistent with the application of the new Map Field rule 3. Assuming the weight that connects unit $J = 0$ to a resonating unit K is less than ρ_{map} (this is true in all cases where lines in the graph have fallen below the black horizontal line, e.g. the 'clock' line) and that which connects L to K is greater than ρ_{map} the rule 3 inhibits ART$_a$ to ART$_b$ Map Field learning.

Following cycle 7 it can be seen from the graphs in figure 7.9, that all of the lines but for one have fallen below the horizontal ρ_{map} line. This means that unit $J = 0$ has established a positive connection with a single unit K since the values of those weights that have fallen below the line cannot again rise. In this case one can see that a strong bond has been established between this unit and the ART$_b$ unit that encodes 'blue'.

Figure 7.10 shows similar zig-zagging early on as $J = 2$ simultaneously resonates with the ART$_b$ units that encode 'orange', 'dog', 'red', and 'cat' during the first 16 resonance cycles. A new phenomena arises in this figure in the guise of the long plateaus that can be seen between cycles 7-10 and 11-15. In these cases rule 3 is being applied over consecutive cycles resulting in no weight changes. However, this can only happen if, in between, rule 4 is being applied, i.e. match tracking is working to stop unit 2 from coding a wrong colour. After 16 cycles unit 2 establishes a strong connection with 'red'.

After zig-zags in relation to 'chair' and 'cat' in cycles 2 and 4, figure 7.11 shows that by the eighth cycle ART$_a$ unit 5 has established a firm connection to the ART$_b$ unit that encode 'green'.

Figure 7.12 again demonstrates the need for match tracking. In early cycles it can be seen that unit 8 is resonating in response to both 'orange' and 'yellow' input patterns. This is not good as the unit will be learning some diluted mix of both colours. Lexical feedback in the form of match tracking is required to force the unit to focus upon only one type of input colour. Due to a greater frequency of yellow objects by cycle 14, the relation to 'orange' has fallen below ρ_{map}. The longer plateaus that follow show that the the unit is again trying to resonate in respect of 'orange' inputs. However, this is being stopped by rule 4. In the

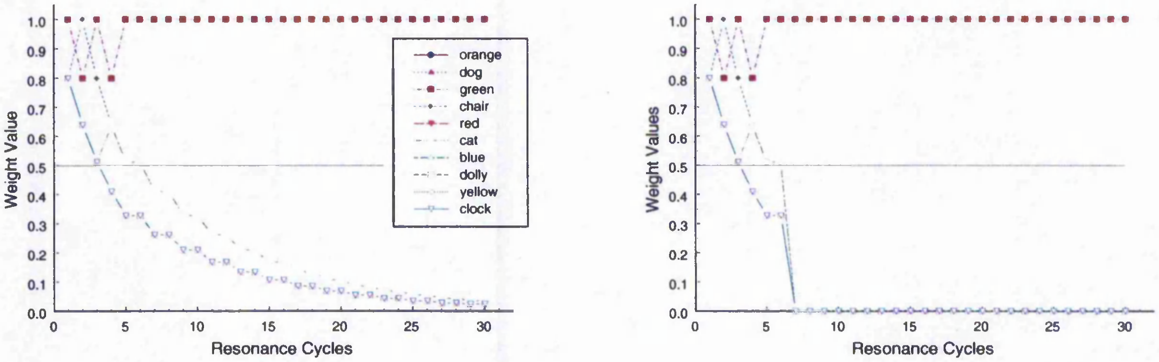

Figure 7.11: The two graphs show how the Map Field weights that connect F_2 unit 5 in ART$_a$ to all F_2 units in ARTb vary over the course of learning.

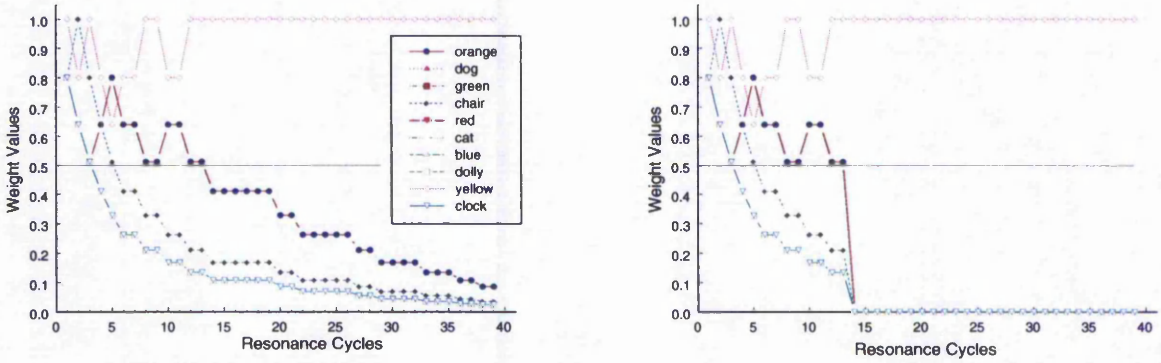

Figure 7.12: The two graphs show how the Map Field weights that connect F_2 unit 8 in ART$_a$ to all F_2 units in ARTb vary over the course of learning.

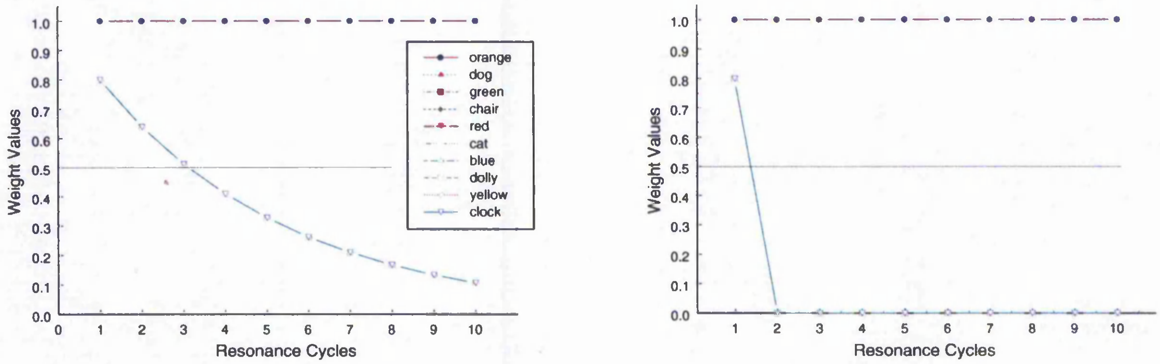

Figure 7.13: The two graphs show how the Map Field weights that connect F_2 unit 9 in ART$_a$ to all F_2 units in ARTb vary over the course of learning.

following cycle rule 3 is keeping weights constant. This rule 3 and 4 interaction again causes this pattern.

Figure 7.13 shows two graphs with interesting variations to those already discussed. No zig-zags are evident with the link to 'orange' staying maximal throughout learning as all others fall during every cycle. This occurs because of the match tracking events described above. 'Orange' input patterns originally cause resonance in ART$_a$ unit 8, however, as discussed above after cycle 14 of unit 8 resonance, the unit has decided it should not be learning about such patterns. The match tracking that stops this causes resonance to occur in another ART$_a$ unit, in this case unit 9. In the following cycle 'orange' again causes resonance first in unit 8, but this time, since the lexical input is a form term, the weight between L and K is greater than ρ_{map} and rule 3 instead of 4 is applied (causing longer plateaus in unit 8). As this happens time after time, the non-orange weights never rise or stay constant but always fall.

The left side graph shows that after 4 cycles unit 9 has established a firm connection with 'orange'. The right side graph shows that by employing fast confident learning this same connection is established after only 2 resonance cycles.

Figures 7.14 through 7.18 show graphs for ART$_c$ resonating units which exhibit patterns similar to those described above. Units 4, 5, 6, 7, and 8 eventually establish connections to 'chair', 'clock', 'dog', 'dolly' and 'cat' respectively.

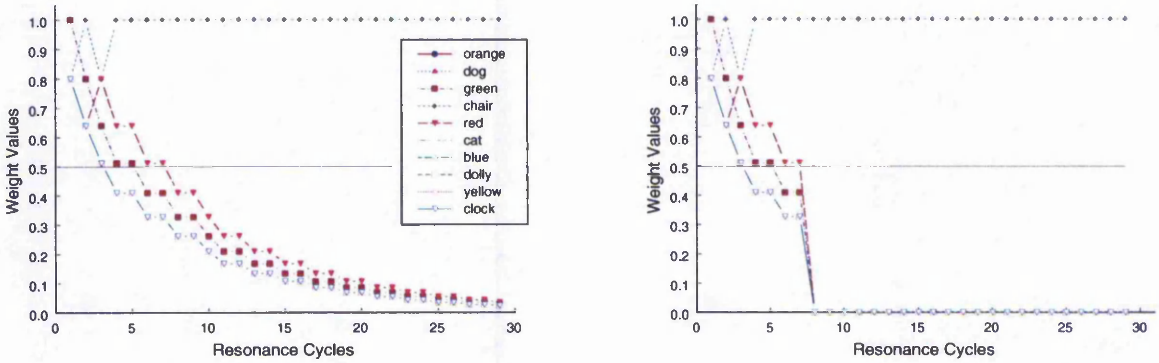

Figure 7.14: The two graphs show how the Map Field weights that connect F_2 unit 4 in ART$_c$ to all F_2 units in ARTb vary over the course of learning.

In order to provide scope for further analysis, a 'snapshot' of the MM model was taken after every 30 learning cycles. At this stage the training process was temporarily halted and the model was tested on the full training set. Thus except for last 'snapshot' this meant the model was actually being tested on some number of unseen coloured objects, i.e. novel data. The task was to describe both the colour and form of the objects by recalling ART$_b$ lexical categories. The graphs presented in figure 7.19 provide the results of these tests.

The left side graph shows how naming of the colour component of the input performs after the various training stages. After 30 training cycles it can be seen that the model is able to name the colour component of green input objects with 100% success. However, naming fails totally for all other coloured objects. This pattern occurs simply because after 30 training cycles only ART$_a$ unit 2 has seen enough data to establish a firm association with a single ART$_b$ category, i.e. only one weight remains above the the ρ_{map} level. As one might expect, performance on the task improves progressively as the amount of training data the model has seen increases. As can be see from the graph it takes the model a lot longer to learn to correctly name orange input objects than objects of any other colour. This is of course because ART$_a$ unit 9 does not start learning about orange till later on (see fig. 7.8) after unit 8 has learnt that it shouldn't be learning about this colour and instead focuses its attention upon learning yellow. After 2 epochs of learning it can be seen that the MM model has learnt to name the colour component of all objects in the training set with complete success.

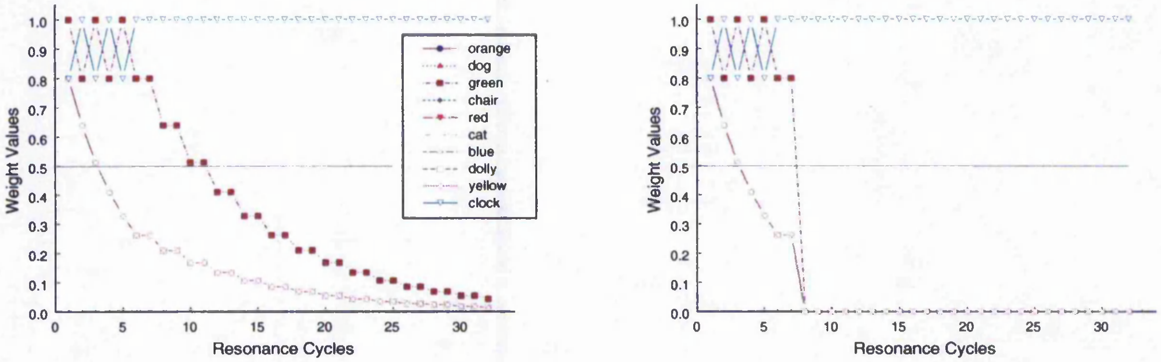

Figure 7.15: The two graphs show how the Map Field weights that connect F_2 unit 5 in ART$_c$ to all F_2 units in ARTb vary over the course of learning.

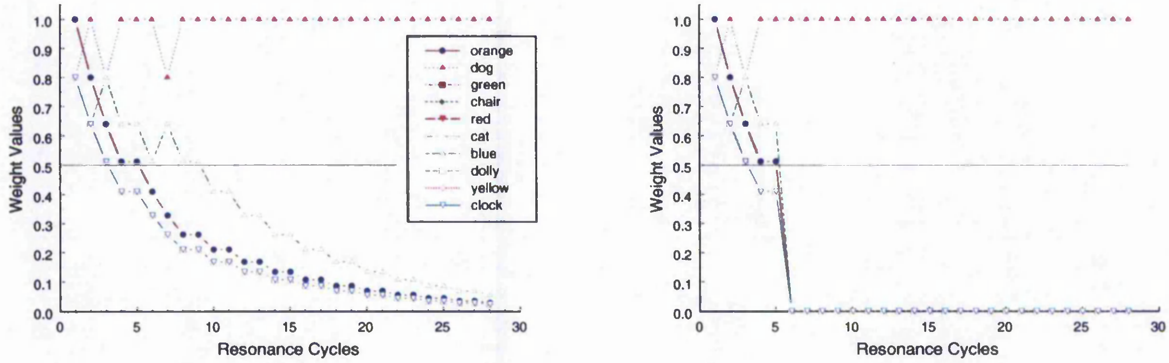

Figure 7.16: The two graphs show how the Map Field weights that connect F_2 unit 6 in ART$_c$ to all F_2 units in ARTb vary over the course of learning.

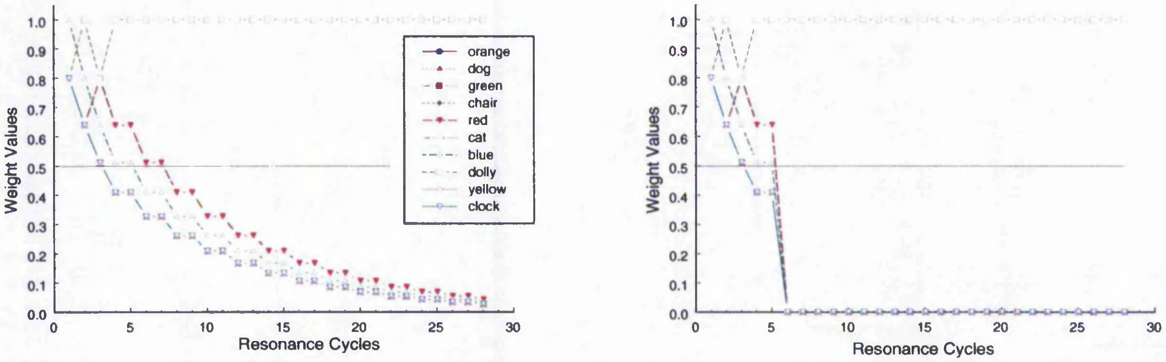

Figure 7.17: The two graphs show how the Map Field weights that connect F_2 unit 7 in ART$_c$ to all F_2 units in ARTb vary over the course of learning.

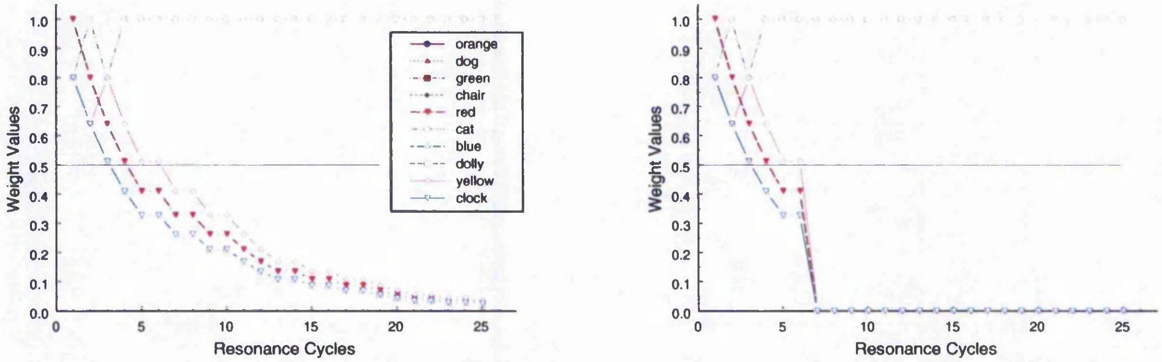

Figure 7.18: The two graphs show how the Map Field weights that connect F_2 unit 8 in ART$_c$ to all F_2 units in ARTb vary over the course of learning.

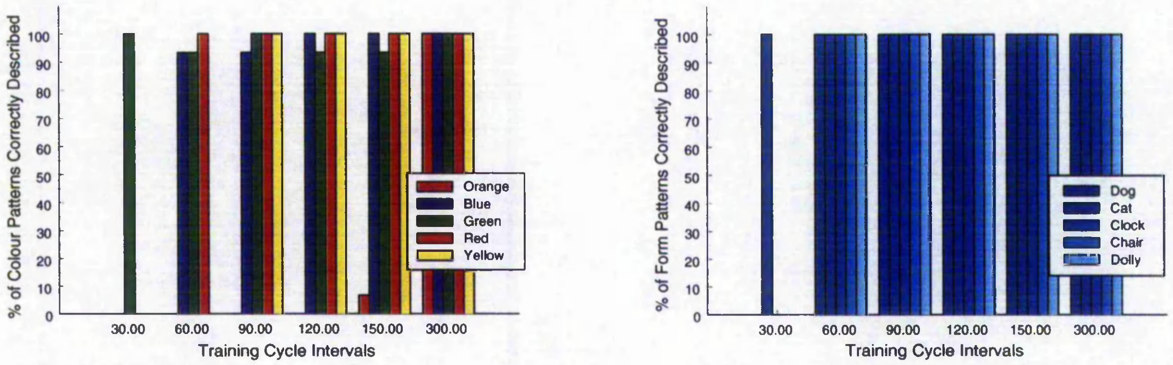

Figure 7.19: The two graphs show how well the model is able to recall, at various snapshot stages, correct lexical categories when provided with coloured objects that make up the training data. The left side graph shows recall of terms resulting from input to the ART$_a$ module. The right side graph shows that resulting from input to the ART$_c$ module.

	Colour Component	Form Component
No. of patterns correctly named	69	75
% of patterns correctly named	92%	100%

Table 7.7: Final results: Shows the number of patterns (out of a total of 75) that were correctly described by the MM model.

The right side graph shows that after 30 training cycles the model is only able to successfully name the form component of cat shaped input objects. However, after 60 cycles it has learnt enough to enable complete success for all objects in the training set.

The final stage of the evaluation process was to test the trained model on a set of novel test objects. The test set comprised a further 75 coloured objects. All of the objects had different shades of the 5 colours to those in the training set. Furthermore, the form component of the data had 10% noise added (as per the experiments in the previous chapter). Table 7.7 shows the results obtained from this testing process.

As can be seen from the table naming of the form component of all of the noisy test objects was completely successful. However, there were a total of 6 errors in correctly naming the colour component of the objects. Further investigation showed that all of the errors arose in relation to the presentation of orange coloured objects and in all cases these were named as yellow objects. The cause of this is down to the fact that in the early stages of training ART$_a$ unit 8, which eventually established a connection with 'yellow', actually learnt a number of orange input patterns before this connection was established. One would expect that further epochs of training, or even better, a much larger training data set containing more examples of each colour would help remedy this problem by giving unit 8 more time to 'forget' its early 'orange' learning as it learns more and more yellow input patterns. Moreover, ART$_a$ unit 9 would be able to learn more orange input patterns.

Overall, the results demonstrate that the new architecture is able to satisfactorily perform MM grounding in respect of two sensory inputs.

Chapter 8

Discussion

The subtitle of this work could be construed as misleading in as much as very little mention has been made of robotic systems throughout the main body of the text. However, whilst the empirical investigations themselves may not have dealt with actual robotic systems it is still felt by the author that further application of this work is best suited to such systems. Moreover, the seeds of the work that has been presented are grounded in investigations that were undertaken with real robotic systems.

In Chandler et al. (1995), this author reported upon a system that had been developed using a commercial automatic speech recognition (ASR) system to drive a robotic measuring device (originally developed by Balendran (1994)). In this system a number of known and distinct states of motor output were mapped to recognized states provided by the ASR system. The limitations of this approach soon became apparent as the system was not able to generalize the mappings it had learnt to other very similar motor states. Each motor to recognition state mapping had to be explicitly defined by the designer. It was clear from this that the meanings of words must surely be more complex than such mappings.

An initial literature review highlighted what have been referred to within this text as "traditional" approaches to addressing such limitations. In AI research early attempts to address meaning and semantics in natural language processing centered upon the design of explicit rules and representations that were said to contain, or capture meaning. Thus, considering the system above, rather than mapping many different states of motor output to a word such as 'fast', one instead employed an explicit rule, e.g. if revolutions per second of the motor is greater than a certain value then the word 'fast' applies. Early AI NLP systems such as SHRDLU (Winograd 1972) relied wholly on such representational formalisms. However, as has been discussed in much greater depth in chapter 2, whilst various forms of such traditional approaches continued to arise so did serious questions that called into disrepute any such systems in which symbolic representations are not grounded. Recent findings (Glenburg & Robertson 2000, Thompson 1997) continue to lend support to such questioning. It was on this basis that further investigations became focussed solely upon grounded language models.

The first grounded model studied in any depth was that described by Dorffner (1992). To gain a better understanding of the model discussed by Dorffner, a computational implementation of his learning equations was undertaken. This implementation, developed in C++ code, helper to uncover the limitations of this particular model (see section 3.2.4 for more detail) and ultimately helped shape a set or requirements that one might expect of a more 'ideal' model (see 3.2.1). A less hands on approach was taken to reviewing other related models as in many cases not enough, if any, detail was provided to facilitate such implementation.

Three main elements were seen to be missing in part or in combination from the reviewed models: plasticity, performance and scalability. In lacking plasticity – the ability to continually learn without corrupting previously learnt information – many of the models reviewed were not able to meet the second objective of this work (as detailed in the opening chapter). It was further discovered that in many cases where plasticity was lacking it was due to the fact that many of the models employed certain kinds of Artificial Neural Networks (ANN) to facilitate learning. These networks were known to lack the ability to learn sequentially (see section 4.1.2). Performance, another of this work's objectives, was also not a major concern for many of the works reviewed. As such whilst one of the more advanced systems DETE (see 3.2.2) offered much by way of direction the lack of consideration for performance meant that the realistic application of such technology was out of the question. Finally, in many models not much consideration was given to the scalability factor. In the context of the objectives of this work such scalability was considered as the ability to ground words with respect to multiple rather than a single sensory domain – thus facilitating a greater breadth of lexicon.

The plasticity factor in the context of ANN learning initially provided the impetus to investigate Adaptive Resonance Theory (ART). The reason being that Grossberg notes that one of the motivating forces (Grossberg 1976*a*, Grossberg 1976*b*) sought to address when developing ART was that of plasticity, suggested that the use of models which were guided by ART's principles may provide a better framework upon which to base a model of lexical acquisition and grounding. To investigate this possibility attention was initially focussed towards a model known as ART1. Based upon detail provided in Carpenter & Grossberg (1987) an implementation of an ART1 model was developed. Again this implementation was created from the ground up using C++ code, the design being led by object oriented principles. Each processing unit in the main ART processing layers (see chapter 4 and appendix A for further explanation) was modeled as an object. This allowed the author to develop an implementation that was able to add new F2 level category units to classify input as and when required in a dynamic fashion. Before application to any LAG related task, the implementation was initially tested against known samples of data (provided in Carpenter & Grossberg (1987)) to ensure that the ART1 dynamics had been correctly implemented.

With a validated ART1 implementation assured, attention was turned to how such a model could acquire a set of colour categories that would be required to ground colour nouns, e.g. 'red', 'blue'. It was on this basis that limitations of the ART1 model first surfaced. On the basis that colour cameras connected to a robot would be able to determine varying levels of RED, GREEN, and BLUE in an input signal it was initially envisaged that such inputs would provide a direct feed into the input layer of the ART1 model. However, the ART1 architecture is unable to take as input analog sources of information such as these but is constrained to processing discrete binary input patterns. Initially, this was not seen as presenting any great problem and a number of methods were devised to encode the three analog variables into binary patterns (see sections 5.2.2 & 5.2.3 [1]). Results obtained from tests with these binary data sets and the ART1 model were disappointing. It was found that the model was not able to generalise well from one colour sample - colour category mapping to another. In many cases it was found that the trained models believed a shade

[1] Whilst in these sections the encodings are discussed in the context of other ART models they were first devised and employed in the early ART1 investigations discussed here.

of colour A to be a much closer relation to a colour B rather than a similar shade of A. Thus colour concepts and/or categories were not being correctly formed.

These findings actually provided an important result. They suggested to the author that the process of arbitrarily encoding sensory representations may be at fault and may actually hamper the ability to learn grounded representations. This suggestion prompted further research which highlighted two other models built upon ART principles, Fuzzy ART and Gaussian ART, both of which were able to process analog inputs.

From here the investigations detailed in chapter 5 started to take shape. Implementations of Fuzzy and Gaussian ART were devised – again following an object oriented methodology – on the basis of the algorithmic detail provided in (Carpenter et al. 1990, Williamson 1995, Williamson 1996) and in the case of Gaussian ART also with help from the architecture's creator[2]. Using known data each of these implementations was validated to ensure accordance with original specifications.

Using the same colour data as discussed in section 5.2 these implementations were put to the test. However, whilst the results were encouraging in comparison to the ART1 results they were still felt to be unsatisfactory. Upon closer inspection it was felt that the unsupervised learning algorithms employed in these models simply lacked enough power to make the task of learning grounded colour categories a simple one (evidence found later in Clark & Thornton (1997) offers validation and a good explanation of why this occurs). This reasoning led the author to discover the ARTMAP architecture – a blueprint for systems that facilitate supervised learning but which still adhere to the basic fundamentals of ART, e.g. the plasticity-stability dilemma.

ARTMAP architectures were also immediately appealing for another reason in that it naturally provided the right sort of framework that could be applied to enable simultaneous acquisition of both lexical and sensory inputs in addition to acquiring the mapping between these. In fact, it provided a good fit with the sort of system discussed in section 3.1.3 (see also figure 3.3). Implementations of Fuzzy and Gaussian ARTMAP devised by this author were put into place and it was these implementations that provided the actual basis for

[2] Many thanks to Mark Williamson (personal communication) for help with implementing and validating the algorithms that make up Gaussian ART and ARTMAP.

the investigations detailed in chapter 5. Given the nature of the ARTMAP architecture it was possible to simply extend upon the coding work that had already been undertaken to implement the unsupervised Fuzzy and Gaussian ART models. Again further validation of the models was required and undertaken.

A further benefit that soon became apparent from the use of ART based algorithms were their inherent performance. The system were able to learn categories at rates of orders of magnitude greater than some of the other LAG models reviewed. It was from these finding that the investigations detailed in chapter 6 were initiated. The implementations devised above were adapted to accommodate the data described in section 6.1.1 and many tests using these were executed on a Pentium 100MHz computing platform having 16Mb of main random access memory. The full empirical results have already been presented in chapter 6. These results provide conclusive evidence that on numerous performance measures ARTMAP architectures excelled. Such performance can be attributed to the nature of the processing employed in these architectures. Localist processing means that learning is greatly focused leading to systems that are required to perform far fewer calculations. ANN researchers have dismissed models that employ such processing citing many pitfalls associated with their use. However, in recent times localist processing is again in favor and a great deal has been done by Page (2000) and Grainger & Jacobs (1998) amongst others to dispell the notion of such pitfalls.

Whilst it was felt that ART based architectures had sufficiently addressed the issues of plasticity and performance, the scalability issue was another matter. Based upon the guiding principles of ART, it was originally envisaged that one may be able to bind three unsupervised ART models in an ARTMAP like fashion making use of the standard ARTMAP algorithms. An implementation was devised upon this basis and the multi-modal grounding task described in chapter 7 was used to assess the applicability of such a model. It was soon evident that something was very wrong with this approach. A long and painstaking analysis of how the architecture performed in relation to 3-way input resulted in the author uncovering the problems detailed in sections 7.2.3 through 7.3.3. Through a process of calculation, trial and error this author developed a number of adaptations to the core learning algorithms employed in the architecture as described in section 7.4. A full implementation was devised and used as the basis of the investigations detailed in the latter half of chapter

7. In terms of the scalability issue a degree of success has certainly been achieved. However, the author also realises that the solution provided may not scale to systems where sensory input is coming from four, five or more pieces of sensory apparatus. Such scope would require further research and assessment.

Chapter 9

Conclusions

In the final chapter of this thesis the major achievements of the work undertaken are presented and then discussed in the context of the aims presented in the opening chapter. These discussions extend to consider possible directions for further work.

The work described in this thesis falls into a number of domains, of which lexical semantics, symbol grounding and adaptive neural networks are considered the most important. In relation to these it is felt that the following contributions have been made:

- A novel graphical means of interpretation has been put forward that seeks to clarify four key levels of representation that should be considered in a symbol grounding context. A clear understanding of these levels within the new Triangle of Meaning has been vital. The empirical investigations in chapter 5 helped to clarify the levels of the new TOM by demonstrating the importance of any grounded models interface with the world, i.e. the nature of its sensory peripheries. Without the basis of understanding that is in essence captured by the new TOM it is felt that this work would have failed in meeting any of its objectives.

- The introduction and application of Adaptive Resonance models to the domains of lexical semantics and symbol grounding. To the best of this authors knowledge such application is novel and has been important in a number of ways that are clarified in a number of the points that follow.

- The introduction of an ART based model in chapter 6 that was able to acquire and ground lexical symbols in relation to visual sensory information. This particular model was shown to address a number of stated aims in that it demonstrated the ability to acquire words and their meanings in an autonomous and incremental fashion. Furthermore, by way of the tests of description and identification it was shown that the meanings the model had acquired were similar to those of humans subjects. Finally, it was shown by way of many empirical tests that the model put forward by this author was able to operate with greater speed and accuracy than comparable models whilst still paying close attention to the efficient use of resources.

- The original development and implementation of a number of adapted learning algorithms and a novel ARTMAP-inspired architecture that fuses three rather than two unsupervised ART modules. This novel system has been applied to the task of lexical acquisition and grounding and has been shown to be an important development in the sense that it has enabled one to ground lexical symbols with respect to more than one single sensory input. This ability was important in the context of meeting the second objective of this work as set out in the first chapter. It should be noted that the adapted learning algorithms and architecture are generic in terms of a learning system and may well have application beyond the context of lexical semantics and symbol grounding. However, such application is beyond the scope of the work detailed here.

However, further to these achievements there are a number of issues that one might raise in relation to the work presented in this thesis from a wider language perspective.

The first point that one might raise in relation to the ART systems discussed in the previous chapters concerns their innate inability to acquire meanings for a large class of lexical symbols. As set out in the introductory chapter the domain of investigation for the work detailed in this thesis was constrained to the creation of a system that could acquire one particular class of lexical symbols, viz. an intrinsically meaningful set of English *nouns.* Although this choice was justified in the opening chapter there was a further reason for selecting this particular category of lexical symbols. Nouns were chosen because the entities to which they typically refer (e.g., people, places, objects) can be thought of as static, i.e. their meanings can be largely derived (as has been shown) from time-independent regularities inherent within sensory information. The referents of most verbs on the other hand have a temporal trait. For example, DETE learnt the meaning of the word 'bounces' by capturing certain temporal relations, i.e. regularities inherent within sequences of visual input (see section 3.2.2). The ART CMs that have been introduced and investigated above lack any temporal processing functionality and therefore cannot capture such regularities. Hence, one may reason correctly, that ART CMs lack the computational machinery required to appropriately ground verbs.

Despite the above, the application of ART *vis-à-vis* lexical semantics is not at a 'dead-end.' Indeed ART based frameworks for temporal learning have been proposed in a number of guises. Hagiwara (Hagiwara 1994) adopts a third F_3 layer of processing units in his Time-Delay ART (TD-ART) system in order to learn the sequence of concept activity at F_2. Carpenter and Grossberg (Carpenter & Grossberg 1994), the originators of ART, have also considered temporal learning and point to ARTSTORE — a computational system that integrates ART CMs with STORE (Sustained Temporal Order REcurrent) models (Bradski et al. 1992, Bradski et al. 1994). The latter capture and then transform time-dependant regularities to a representational format that can be learnt and recalled using ART CMs. Moreover, it has been shown by way of mathematical, computational, and simulation data (Grossberg & Merrill 1996) that ARTSTORE systems escape an obstacle that plagues many alternative temporal learning architectures, namely the inability to acquire large amounts of temporally ordered information rapidly and in a stable fashion, i.e. in the same way that ART models solve the stability-plasticity dilemma in relation to spatially defined input data, ARTSTORE systems do so for temporally defined data.

Thus it is suggested that such systems could be harnessed to provide a framework that supports the learning of verbs and their associated meanings. This is an avenue for further investigation.

Another issue that has been addressed in related research (see Dorffner's work; section 3.2.4) but has been neglected here is that of lexical taxonomy. Simply put, evidence suggests that the world is not conceptualized at a single level, but rather that there is a multi-level hierarchy consisting of subordinate level categories (e.g. poodle), basic level categories (e.g. dog), and superordinate level categories (e.g. animal). Here again ART is unique in connectionist terms in that one has direct control over the degree of generalization through variation of the vigilance parameter. Bartfai (Bartfai 1995) has already created such ART-based models called HART (Hierarchical ART) which employ multiple values of vigilance in order to achieve multi-level pattern clustering. Again, such models provide an avenue for further investigation.

Another issue worthy of note in such a discussion is that that of linguistic naivity. Many who practice the methods of traditional semantic theory will argue that models such as those presented herein have a very limited vocabulary and are thus lexically naive (for evidence of this refer to (Dorffner 1992, Chrisley 1994)). It is conceded that the breadth of the vocabularies used in those models discussed in chapter 3 and in the new ART-based models introduced in the latter chapters are limited. However, 'naivity' is contended, since these models have a much greater depth of lexical understanding for those words that they do use. The principle followed has been that one must take a step back in order to take two in a forward direction. Many traditional language models, whilst having a large vocabulary, have proved to be limited in their domain of application because what they know of the words they use is very constrained. It is hoped that by using a grounded vocabulary in the future new language models built upon the principles discussed in this thesis will not run up against the barriers that have stopped the traditional models in their tracks.

The investigations and discussions in chapter 5 noted important ramifications for lexical acquisition and grounding models *vis-à-vis* how their sensory peripheries measure aspects of the external world. In this context it is suggested that LAG models will improve as science uncovers more about the unconscious and cognitively impenetrable conceptual skills that humans take for granted. Moreover, improvements should also come, as such systems are able to interact with the world (e.g. via artificial limbs) with greater ease and dexterity, e.g. see Saffiotti & LeBlanc (2000). It has been suggested that a lack of such advanced interaction is something that has held back the development of grounded language models so far (MacDorman 1999, Ziemke 1999).

To conclude it is felt that the work that has been described herein has been successful in meeting the aim and objectives originally conceived. However, in a wider context and as the title of this thesis suggests in its use of the word 'towards', there is still much to be done to create grounded language systems are truly useful. The previous couple of paragraphs suggest possible paths forward and it is hoped that others will travel this path and that the work and discussions provided within thesis have provided at least some inspiration to do so. Moreover, it is hoped that language models based upon some of the principles introduced in this thesis will one day provide artificial agents (e.g. robotic systems) with an enhanced and powerful means by which they can communicate with the human species.

References

Allen, J. (1995), *Natural Language Understanding*, Addison-Wesley.

Balendran, V. (1994), Cosmetic Quality of Surfaces: A Computational Approach, PhD thesis, Nottingham Trent University.

Bartell, B. & Cottrell, G. (1991), A model of symbol grounding in a temporal environment, *in* 'AAAI Spring Symposium Workshop on Connectionist Natural Language Processing'.

Bartfai, G. (1995), An art-based modular achitecture for learning hierachical clusterings, Technical Report CS-TR-95/3, Victoria University of Wellington, New Zealand.

Bates, M. & Weischedel, R., eds (1993), *Challenges in Natural Language Processing (Studies in Natural Language Processing)*, Cambridge University Press.

Berlin, B. & Kay, P. (1969), *Basic Color Terms: Their Universality and Evolution*, University of California Press, Berkeley CA.

Bickhard, M. & Terveen, L. (1995), *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution*, Vol. 109 of *Advances in Psychology*, Elsevier Science.

Bradski, G., Carpenter, G. & Grossberg, S. (1992), Working memories for storage and recall of arbitrary temporal sequences, *in* 'Proceedings of the International Joint Conference on Neural Networks (IJCNN-92)', IEEE, Piscataway, NJ, pp. 57–62.

Bradski, G., Carpenter, G. & Grossberg, S. (1994), 'Store working memory networks for storage and recall of arbitrary temporal sequences', *Biological Cybernetics* **71**, 469–480.

Carnap, R. (1952), 'Meaning postulates', *Philosophical Studies* **3**, 65–73.

Carpenter, G. & Grossberg, S. (1987), 'A massively parallel architecture for a self-organising neural pattern recognition machine', *Computer Vision, Graphics, and Image Processing* **37**, 54–115.

Carpenter, G. & Grossberg, S. (1988), 'The art of adaptive pattern recognition by a self-organizing neural network', *Computer* **21**(3), 77–88.

Carpenter, G. & Grossberg, S. (1994), Integrating symbolic and neural processing in a self-organising architecture for pattern recognition and prediction, *in* V. Honavar & L. Uhr, eds, 'Artificial Intelligence and Neural Networks: Steps toward Principled Integration', Academic Press, Boston, MA, pp. 387–420.

Carpenter, G., Gjaja, M., Gopal, S., Markuzon, N. & Woodcock, C. (1996), Art and artmap neural networks for applications: Self-organizing, learning, recognition, and prediction, Technical Report CAS/CNS-96-009, Boston University, Boston, MA.

Carpenter, G., Grossberg, S. & Reynolds, J. (1991), 'Artmap: Supervised real-time learning and classification of nonstationary data by a self-organising neural network', *Neural Networks* **4**, 565–588.

Carpenter, G., Grossberg, S. & Rosen, D. (1990), 'Fuzzy art: Fast stable learning and categorisation of analog patterns by an adaptive resonance system', *Neural Networks* **4**, 759–771.

Carpenter, G., Grossberg, S., Markuzon, N., Reynolds, J. & Rosen, D. (1992), 'Fuzzy artmap: A network architecture for incremental supervised learning of analog multidimensional maps', *IEEE Transactions on Neural Networks* **3**, 698–713.

Caudell, T., Smith, S., Escobedo, R. & Anderson, M. (1994), 'Nirs: Large scale art 1 neural architectures for engineering design retreival', *Neural Networks* **7**, 1339–1350.

Chalmers, D. (1992), Subsymbolic computation and the chinese room, *in* J. Dinsmore, ed., 'The symbolic and connectionist paradigms: Closing the gap', Lawrence Erlbaum, Hillsdale, NJ.

Chandler, N., Balendran, V., Evett, L. & Sivayoganathan, K. (1995), From speech recognition to language understanding: An engineering perspective, *in* '11th National Conference on Manufacturing Research', DeMontfort University, England, pp. 562–566.

Chandler, N., Balendran, V., Evett, L. & Sivayoganathan, K. (1999), Reality: A prerequisite to meaningful representations, *in* A. Riegler & M. Peschl, eds, 'Understanding Representation in the Cognitive Sciences: Does Representation Need Reality?', Kluwer Academic/Plenum Publishers, NY.

Chauvin, Y. (1989), Toward a connectionist model of symbol emergence, *in* 'Proceedings of the Eleventh Annual Conference of the Cognitive Science Society', Erlbaum, Hillsdale, NJ, pp. 580–587.

Chierchia, G. & McConnell-Ginet, S. (1990), *Meaning and Grammar: An Introduction to Semantics*, MIT Press, Cambridge, MA.

Chomsky, N. (1980), 'Rules and representations', *Behavioral and Brain Sciences* **3**, 1–61.

Chrisley, R. (1994), 'Conceptulizing how dete conceptualizes (or, 'more dete-tales, please!')', *Connection Science* **6**(1), 113–115.

Clark, A. & Thornton, C. (1997), 'Trading spaces: Computation, representation and the limits of uninformed learning', *Behavioral and Brain Sciences* **20**(1), 57–92.

Clark, E. (1973), What's in a word? on the child's acquisition of semantics in his first language, *in* T. Moore, ed., 'Cognitive Development and the Development of Language', Academic Press.

Cybenko, G. (1989), 'Approximation by superpositions of a sigmoidal function', *Mathematics of Control, Signals, and Systems* **2**, 303–314.

Dennett, D. (1969), *Content and Consciousness*, Routlage and Kegan Paul, London.

Dennett, D. (1997), When hal kills, who's to blame? computer ethics, *in* D. Stork, ed., 'Hal's Legacy: 2001's Computer as Dream and Reality', MIT Press.

Dorffner, G. (1992), A step toward sub-symbolic language models without linguistic representations, *in* R. Reilly & N. Sharkey, eds, 'Connectionist Approaches to Natural Language Processing', Earlsdale, pp. 273–313.

Dorffner, G. (1994), ‘Why connectionism and language modelling need dete’, *Connection Science* **6**(1), 115–118.

Dorffner, G. (1995), On grounding language with neural networks, *in* ‘IEE Colloquium on Grounding Representations: Integration of sensory information in natural language processing, artificial intelligence, and neural networks’, IEE, London.

Dorffner, G. (1997), The connectionist route to embodiment and dynamicism, *in* A. Riegler & M. Peschl, eds, ‘Does Representation need Reality? - Proceedings of the International Conference ‘New Trends in Cognitive Science”, Austrian Society for Cognitive Science Technical Report 97-01, Vienna, Austria.

Dubrawski, A. & Crowley, J. (1994), ‘Learning locomotion reflexes: A self-supervised neural system for a mobile robot’, *Robotics and Autonomous Systems* **12**, 133–142.

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

Feldman, J., Lakoff, G., Bailey, D., Narayanan, S., Reiger, T. & Stolcke, A. (1996), ‘L_0 - the first five years of an automated language acquisition project’, *Artificial Intelligence Review* **10**, 103–129.

Fodor, J. A. (1975), *The Language of Thought*, Harvard University Press, Cambridge, MA.

Fodor, J. A. (1980), ‘Searle on what only brains can do’, *The Behavioral and Brain Sciences* **3**, 431–432.

French, R. (1997), Pseudo-recurrent connectionist networks and the problem of sequential learning, *in* ‘Forth Neural Computation and Psychology Workshop; Connectionist Representations : Theory and Practice’, University of London, England.

Gärdenfors, P. (1995), Meanings as conceptual structures, Technical Report LUCS 40, Lund University Cognitive Science Department.

Gasser, M. & Smith, L. (1993), Learning noun and adjective meanings: A connectionist account, Technical report, Department of Computer Science and Linguistics, Indiana University. technical report.

Gazdar, G. (1993), The handling of natural language, *in* D. Broadbent, ed., ‘The Simulation of Human Intelligence’, Blackwell, Oxford, pp. 151–177.

Gazdar, G. & Mellish, C. (1989), *Natural Language Processing in PROLOG: an Introduction to Computational Linguistics*, Addison-Wesley.

Glenburg, A. (1997), 'What memory is for', *Behavioral and Brain Sciences* **20**(1), 1–55.

Glenburg, A. & Robertson, D. A. (2000), 'Symbol grounding and meaning: A comparison of highdimensional and embodied theories of meaning', *Journal of Memory and Language* **43**, 379–401.

Gorin, A. (1995), 'On automated language acquisition', *Journal of the Acoustical Society of America* **97**(6), 3441–3461.

Grainger, J. & Jacobs, A. (1998), 'Localist connectionism fits the bill', *Psycoloquy*. ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1998.volume.9/psyc.98.9.10.connectionist-explanation.7.grainger.

Grishman, R. (1986), *Computational Linguistics: an introduction - (Studies in natural language processing)*, Cambridge University Press.

Grossberg, S. (1976*a*), 'Adaptive pattern clasification and universal recoding, i: Parallel development and coding of neural feature detectors', *Biological Cybernetics* **23**, 121–134.

Grossberg, S. (1976*b*), 'Adaptive pattern clasification and universal recoding, ii: Feedback, expectation, olfaction, and illusions', *Biological Cybernetics* **23**, 187–202.

Grossberg, S. (1987), 'Competitive learning: From interactive activation to adaptive resonance', *Cognitive Science* **11**, 23–63.

Grossberg, S. (1988), 'Nonlinear neural networks: Principles, mechanisms, and architectures', *Neural Networks* **1**, 17–61.

Grossberg, S. (1997), 'Neural models of development and learning', *Behavioral and Brain Sciences* **20**(4), 566.

Grossberg, S. & Merrill, W. (1996), 'The hippocampus and cerebellum in adaptively timed learning, recognition, and movement', *Journal of Cognitive Neuroscience* **8**, 257–277.

Grumbach, A. (1996), 'Grounding symbols into perceptions', *Artificial Intelligence Review* **10**, 131–146.

Hagiwara, M. (1994), 'Time-delay art for spatio-temporal patterns', *Neurocomputing* **6**, 513–521.

Hanson, S. & Burr, D. (1990), 'What connectionist models learn: Learning and representation in connectionist networks', *Behavioral and Brain Sciences* **13**, 471–518.

Harnad, S. (1990), 'The symbol grounding problem', *Physica D* **42**(1/3), 335–346.

Harnad, S., Hanson, S. & Lubin, J. (1994), Learned categorical perception in neural nets: Implications for symbol grounding, *in* V. Honvar & L. Uhr, eds, 'Artificial Intelligence and Neural Networks: Steps Toward Principled Integration', Academic Press, Boston, MA.

Henis, E. & Levinson, S. (1995), Language as part of sensorimotor behavior, *in* 'AAAI Fall Symposium on Embodied Langauge and Action', pp. 48–53.

Hirst, G. (1987), Semantics, *in* S. Shapiro, ed., 'Encyclopedia of Artificial Intelligence', Wiley, New York, pp. 1024–1029.

Hornik, K., Stinhcombe, M. & White, H. (1989), 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**, 359–366.

Howarth, M. (1998), An Investigation of Task Level Programming for Robotic Assembly, PhD thesis, Nottingham Trent University.

Jackson, S. & Sharkey, N. (1996), 'Grounding computational engines', *Artificial Intelligence Review* **10**, 65–82.

Jurafsky, D. & Martin, J. (2000), *Speech and Language Processing*, Prentice Hall, New Jersey.A

Karlsson, F. & Karttunen, L. (1996), Sub-sentential processing, *in* R. Cole, ed., 'Survey of the State of the Art in Human Language Technology'. www.cse.ogi.edu/CSLU/HLTsurvey/.

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

Katz, J. & Fodor, J. (1963), 'The structure of a semantic theory', *Language* **39**, 170–210.

Keat, J. (1996), Adaptive Invariant Recognition and Assessment of Free-Form Objects: A Connectionist Approach, PhD thesis, Nottingham Trent University.

Kohonen, T. (1982), 'Self-organized formation of topologically correct feature maps', *Biological Cybernetics* **43**, 59–69.

Lakoff, G. (1987), *Women, Fire, and Dangerous Things; What Categories Reveal about the Mind*, University of Chicago Press.

Lammens, J. (1994), A Computational Model of Colour Perception and Colour Naming, PhD thesis, State University of New York.

Landau, B. & Jackendoff, R. (1993), '"what" and "where" in spacial language and spacial cognition', *Behavioral and Brain Sciences* **16**(2), 217–265.

Lenat, D. (1990), *Building Large Knowledge-based Systems: Representation and Inference in the CYC Project*, Addison-Wesley.

Lenat, D. (1997), From 2001 to *2001*: Common sense and the mind of hal, *in* D. Stork, ed., 'Hal's Legacy: 2001's Computer as Dream and Reality', MIT Press.

Livingstone, M. & Hubel, D. (1988), 'Segregation of form, colour, movement, and depth: Anatomy, physiology, and perception', *Science* **240**, 740–749.

Lyons, J. (1977), *Semantics*, Cambridge University Press.

MacDorman, K. (1995), How to ground symbols adaptively, *in* 'AISB-95 Workshop Reaching for Mind: Foundations of Cognitive Science'.

MacDorman, K. (1999), 'Grounding symbols through sensorimotor integration', *Journal of the Robotics Society of Japan* **17**(1), 20–24.

Marconi, D. (1996*a*), 'On the referential competence of some machines', *Artificial Intelligence Review* **10**, 21–35.

Marconi, D. (1996*b*), 'Work on the integration of language and vision at the university of torino', *Artificial Intelligence Review* **10**, 15–20.

Marr, D. (1982), *Vision*, Freeman, New York.

Maturana, H. & Varela, F. (1980), *Autopoiesis and Cognition: The Realization of the Living*, D. Reidel, Dordrecht.

McClelland, J. & Rumelhart, D. (1981), 'An interactive activation model of context effects in letter perception', *Psychological Review* **88**, 375–407.

McCloskey, M. & Cohen, N. (1989), Catastrophic interference in connectionist networks: The sequential learning problem, *in* G. Bower, ed., 'The Psychology of Learning and Motivation', Academic Press, New York, pp. 109–165.

McDermott, D. (1976), 'Artificial intelligence meets natural stupidity', *SIGART* **57**, 4–9.

McShane, J. (1979), 'The development of naming', *Linguistics* **17**, 879–905.

Montague, R. (1988), The proper treatment of quantification in ordinary english, *in* J. Kulas, J. H. Fetzer & T. L. Rankin, eds, 'Philosophy, Language, and Artificial Intelligence: Resources for Processing Natural Language', Kluwer, Boston, pp. 141–162.

Mooney, R. (1985), Generalising explanations of narratives into schemata, Technical Report T147, CSL: University of Illinois, Urbana.

Nenov, V. & Dyer, M. (1988), Dete: Connectionist-symbolic model of visual and verbal association, *in* 'IEEE International Conference on Neural Networks', Vol. 2, pp. 17–24.

Nenov, V. & Dyer, M. (1993), 'Perceptually grounded language learning: Part i - a neural network architecture for robust sequence association', *Connection Science* **5**(2), 115–138.

Nenov, V. & Dyer, M. (1994), 'Perceptually grounded language learning: Part ii - a neural/procedural model', *Connection Science* **6**(1), 3–41.

Ogden, C. & Richards, I. (1923), *The Meaning of Meaning*, Routledge and Kegan Paul, London.

Page, M. (2000), 'Connectionist modelling in psychology: A localist manifesto', *Behavioral and Brain Sciences.*

Partee, B. (1981), Montague grammar, mental representations, and reality, *in* S. Oehman & S. Kanger, eds, 'Philosophy and Grammar', Dordrecht, pp. 59–78.

Peschl, M. (1992), 'Construction, representation, and the embodiment of knowledge, meaning, and symbols in neural structures: Towards an alternative understanding of knowledge representation and philosophy of science', *Connection Science* **4**(3/4), 327–338.

Pinker, S. (1994), *The Language Instinct*, Penguin Books, London.

Plunkett, K., Sinha, C., Møller, M. & Strandsby, O. (1992), 'Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net', *Connection Science* **4**(3/4), 293–312.

Prem, E. (1994), Symbol grounding revisited, Technical Report OEFAI-TR-94-19, Österreichisches Forschungsinstitut für Artificial Intelligence. ftp://ftp.ai.univie.ac.at/papers/oefai-tr-94-19.ps.z.

Putnam, H. (1981), *Reason, Truth, and History*, Cambridge University Press.

Pylyshyn, Z. (1980), 'The causal power of machines', *Behavioral and Brain Sciences* **3**, 442–444.

Pylyshyn, Z. (1987), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Ablex, Norwood, NJ.

Quartz, S. & Sejnowski, T. (1997), 'The neural basis of cognitive development: A constructivist manifesto', *Behavioral and Brain Sciences* **20**, 537–596.

Rorty, R. (1980), *Philosophy and the Mirror of Nature*, Basil Blackwell, Oxford.

Rosch, E. (1973), 'Natural categories', *Cognitive Psychology* **4**, 328–350.

Rumelhart, D. & Zipser, D. (1985), 'Feature discovery by competitive learning', *Cognitive Science* **9**, 75–112.

Rumelhart, D., McClelland, J. & the PDP Research Group, eds (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol.1: Foundations*, Bradford Books/MIT Press.

Sabah, G. (1993), 'Knowledge representation and natural language understanding', *AICOM* **6**(3/4), 155–186.

Saffiotti, A. & LeBlanc, K. (2000), Active perceptual anchoring of robot behavior in a dynamic environment, *in* 'Proc. of the IEEE Int. Conf. on Robotics and Automation', San Francisco, CA.

Sales, N. (1996), Aspects of Lexical Development in Artificial Neural Networks, PhD thesis, Imperial College, London.

Schank, R. (1973), Identification of conceptualizations underlying natural language, *in* R. Schank & K. Colby, eds, 'Computer models of thought and language', Freeman, San Francisco, pp. 187–247.

Schank, R. & Abelson, R. (1977), *Scripts, Plans, Goals and Understanding*, Lawrence Erlbaum, Hillsdale, New Jersey.

Schyns, P. & Rodet, L. (1995), Concept learning, *in* M. Arbib, ed., 'The Handbook of Brain Theory and Neural Networks', MIT Press, Cambridge, MA, pp. 234–238.

Searle, J. (1980), 'Minds, brains, and programs', *Behavioral and Brain Sciences* **3**, 414–457.

Seibert, M. & Waxman, A. (1992), 'Adaptive 3d object recognition from multiple views', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**, 107–124.

Shapiro, S. C. (2000), SNePS: A logic for natural language understanding and commonsense reasoning, *in* L. Iwańska & S. C. Shapiro, eds, 'Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language', AAAI Press/The MIT Press, Menlo Park, CA, pp. 175–195.

Sharkey, N. & Jackson, S. (1994), Three horns of the representational trilemma, *in* V. Honvar & L. Uhr, eds, 'Symbol Processing and Connectionist Models for Artificial Intelligence and Cognitive Modelling: Steps Towards Integration', Academic Press, pp. 155–189.

Sharkey, N. & Sharkey, A. (1995), 'An analysis of catastrophic interference', *Connection Science* **7**(3/4), 301–329.

Sowa, J. (1984), *Conceptual Structures: Information Processing in Mind and Machine. The systems programming series*, Addison-Wesley, Reading, MA.

Sowa, J. (1987), Semantic networks, *in* S. Shapiro, ed., 'Encyclopedia of Artificial Intelligence', Wiley, New York, pp. 1011–1024.

Sowa, J. (1991), *Principles of Semantic Networks: Explorations in the Representation of Knowledge*, Morgan Kaufman, San Mateo, CA.

Thompson, E. (1997), 'Symbol grounding: A bridge from artificial life to artiificial intelligence', *Brain and Cognition* **34**, 48–71.

Thorpe, S. (1995), Localized versus distributed representations, *in* M. Arib, ed., 'The Handbook of Brain Theory and Neural Networks', MIT Press, Cambridge, MA, pp. 549–552.

Turing, A. (1950), 'Computing machines and intelligence', *Mind* **59**, 433–460.

van Gelder, T. (1991), Classical questions, radical answers: Connectionism and the structure of mental representations, *in* T. Horgan & J. Tienson, eds, 'Connectionism and the Philosophy of Mind', Klewer, Dordrecht.

Wilks, Y. (1975), 'A preferential, pattern-seeking semantics for natural language inference', *Artificial Intelligence* **6**, 53–74.

Williamson, J. (1995), Gaussian artmap: A neural network for fast incremental learning of noisy multidimensional maps, Technical Report CAS/CNS-95-003, Dept. of Cognitive and Neural Systems, Boston University, Boston, MA.

Williamson, J. (1996), A constructive incremental-learning network for mixture modelling and classification, Technical Report CAS/CNS TR-96-019, Dept. of Cognitive and Neural Systems, Boston University, Boston, MA.

Winograd, T. (1972), *Understanding Natural Language*, Academic Press, New York.

Winograd, T. (1973), A procedural model of language understanding, *in* R. Schank & K. Colby, eds, 'Computer Models of Thought and Language', Freeman, San Francisco.

Winograd, T. (1980), 'What does it mean to understand language?', *Cognitive Science* **4**, 209–241.

Wittgenstein, L. (1953), *Philosophical Investigations*, Macmillan, New York.

Woods, W. (1968), Procedural semantics for a question-answering machine, *in* 'AFIPS Conf. Proc 33 (Fall Joint Computer Conference)', pp. 457–471.

Ziemke, T. (1999), Rethinking grounding, *in* A. Riegler & M. Peschl, eds, 'Understanding Representation in the Cognitive Sciences: Does Representation Need Reality?', Kluwer Academic/Plenum Publishers, New York.

Appendices

Appendix A

ART Model Dynamics

The work described in the main body of this thesis has employed and extended upon two particular ART based CMs: Fuzzy ARTMAP, and Gaussian ARTMAP. As a point of reference this appendix details an algorithmic description of these models, and of the relevant unsupervised ART models upon which they rely, in a step-by-step fashion.

A.1 Fuzzy ART

A.1.1 Component Specification

The input **a** to a Fuzzy ART architecture is an M-dimensional vector ($\{a_1, \ldots, a_M\}$), where each element has a value in the interval [0,1]. It follows that the F_0 layer of the ART system must have M units. In an ART system that employs complement coding (see below) as a means of input normalization F_1 must contain twice as many units as F_0, i.e. $2M$. Finally, the F_2 layer contains N units. Both M, and N may be arbitrariliy large.

Each F_2 level unit j (where $j = 1, \ldots, N$) has an associated weight vector w_j (a single vector suffices because in a fuzzy ART model the bottom-up LTM weights, and the top-down LTM weights are functionally equivalent, having the same values at all times) that has $2M$ weights, each connecting the unit j to a different F_1 unit i (where $i = 1, \ldots, 2M$).

All weights are initialized with a unitary value:

$$w_{j1} = \ldots = w_{j2M} = 1.0 \tag{A.1}$$

and each category is said to be uncommitted.

A.1.2 Model Parameters

The dynamics of Fuzzy ART CMs are controlled by way of three parameters:

1. A Choice Parameter, α: The value of α must be > 1.0. If α is small, categories with small weight vectors $|w_j|$, and thus occupying large regions of feature space, are favoured; if α is large, the opposite.

2. A Learning Rate Parameter, β: The value of β is set in the interval [0,1]. Employing the maximum value will result in fast one-step learning. With lower values, the adaptation of weights is undertaken with greater moderation. This can provide greater resiliance to noise.

3. A Vigilance Parameter, ρ: The value of ρ is set in the interval [0,1]. Low vigilance effects broad generalization, i.e. categories that occupy large regions of feature space encompassing many possible input patterns. Conversly, high vigilance effects less generalization creating smaller categorical regions.

A.1.3 Processing functions

Input Preprocessing

Complement coding is a preprocessing step that normalizes a raw input vector (see (Carpenter et al. 1990) for a detailed discussion). The preprocessed input vector I that is output from F_0 is twice the size of the raw input vector **a**, where

$$I = (a, a^c)$$

and $\mathbf{a}^c$ is the complement of $\mathbf{a}$ (i.e. $\{1 - a_1, \ldots, 1 - a_M\}$). Thus if the M-dimensional vector

$$\mathbf{a} = \{0, 1, 1\}$$

then $\mathbf{a}^c = \{(1-0),(1-1),(1-1)\}$ and thus the $2M$-dimensional vector I is as follows:

$$I = \{0,1,1,1,0,0\}$$

Bottom-up Activation and F_2 Choice

Each F_2 category unit j aggregates the signals it receives by way of the following fan-in (activation) function

$$T_j = \frac{|I \wedge w_j|}{\alpha + |w_j|} \tag{A.2}$$

where the fuzzy AND operator $\wedge$ chooses the lower of two values, e.g.

$$(0.4 \wedge 0.7) = 0.4$$

and where $|.|$ denotes the magnitude of a vector, which is measured by summing the values of a vectors elements, e.g.

$$|w_j| \equiv \sum_{i=0}^{2M} w_{ji}$$

Competition is engaged at the F_2 layer by way of a choice function that selects the F_2 unit J that has maximal $\mathbf{T}$ activity

$$T_J = max\{T_j : j = 1, \ldots, N\} \tag{A.3}$$

Top-down Hypothesis Testing

By choosing unit J an ART system makes a hypothesis about which category I belongs to. The category representation connected with this category is that stored in the weights which connect unit J to all F_1 units, i.e. w_J. The matching function $(|I \wedge w_J|)(|I|)^{-1}$ measures how close this representation is to the current input vector. If this measurement is small enough in relation the systems vigilance measure; that is, if

$$\frac{|I \wedge w_J|}{|I|} \geq \rho \tag{A.4}$$

then I and w_J are deemed to be close enough (a match) and learning is triggered (see below), however, if

$$\frac{|I \wedge w_J|}{|I|} < \rho \tag{A.5}$$

then J is a bad hypothesis. In this case T_J is set to -1 (reset) in order that unit J can no longer compete to code I. A Search process is initiated whereby a new J is chosen, by equation A.3. The search continues until a J is chosen where equation A.4 is satisfied.

Learning

Once search ends, the weight w_J is updated according to the equation

$$w_J^{(new)} = \beta(I \wedge w_J^{(old)}) + (1 - \beta)w_J^{(old)} \tag{A.6}$$

The Fuzzy ART processing cycle described above is depicted graphically in figure A.1

A.2 Gaussian ART

A.2.1 Component Specification

The physical specification of a Gaussian ART CM differs from its fuzzy counterpart in two ways. Firstly, in a Gaussian CM there is no need to employ complement coding and as such the F_1 layer contains the same number of units M as the F_0 layer. Secondly, each F_2 unit is linked to the units in the F_1 layer by way of *two* weight vectors μ_j (which controls the mean position of the Gaussian category, i.e. boundary in feature space) and σ_j (which controls the variance, i.e. width of the category). In addition each Gaussian F_2 unit has an associated parameter n_j, which measures the number of input patterns a particular unit has coded, and thus is initialized with a value of zero. The other parameters are initialized as follows

$$\mu_{j1} \quad = \ldots = \mu_{jM} \quad = 1.0 \tag{A.7}$$

$$\sigma_{ji} \quad = \ldots = \sigma_{jM} \quad = \gamma \tag{A.8}$$

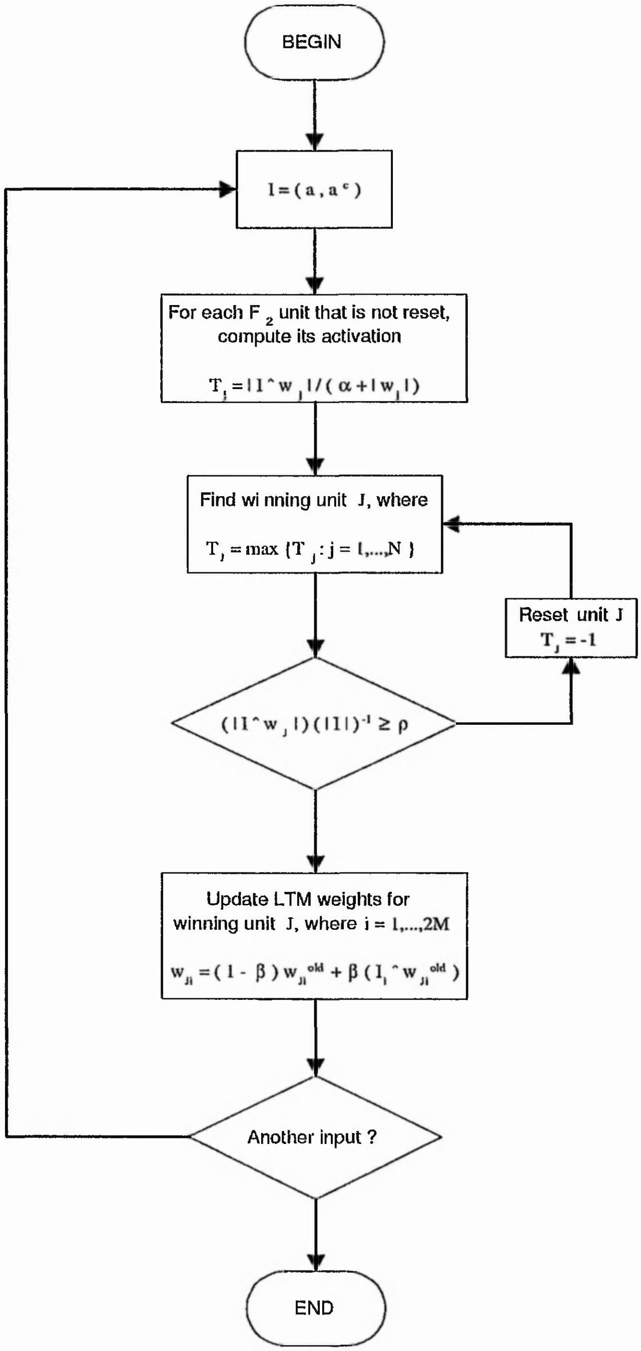
BEGIN
$I = (a, a^c)$
For each F_2 unit that is not reset, compute its activation
$T_j = |I \wedge w_j| / (\alpha + |w_j|)$
Find wi nning unit J, where
$T_J = \max\{T_j : j = 1,...,N\}$
Reset unit J
$T_J = -1$
$(|I \wedge w_J|)(|I|)^{-1} \geq \rho$
Update LTM weights for winning unit J, where i = 1,...,2M
$w_{Ji} = (1 - \beta) w_{Ji}^{old} + \beta (I_i \wedge w_{Ji}^{old})$
Another input ?
END

Figure A.1: Figure shows the algorithmic processing cycle employed in the implementation of an unsupervised Fuzzy ART CM.

A.2.2 Model Parameters

The dynamics of Gaussian ART CMs are controlled by way of two parameters:

1. A Vigilance Parameter, ρ: (see section A.1.2 above).

2. A Variance Initialization Parameter, γ: This parameter is used to initialize the variance vector σ_j for each Gaussian category unit (see eq. A.8). Large values facilitate slower, more graceful noise tolerant learning, and vice versa.

A.2.3 Processing functions

Input Preprocessing

There is no requirement for input preprocessing in Gaussian CMs, however, the architecture performs best when the values of each element of the input vectors it processes have similar bounds.

Bottom-up Activation and F_2 Choice

Each F_2 category unit j aggregates the signals it receives by way of the following fan-in (activation) function

$$T_j = \frac{n_j \, G_j}{\sigma_j} \tag{A.9}$$

where

$$G_j = \exp\left(-\frac{1}{2}\left(\frac{I - \mu_j}{\sigma_j}\right)^2\right) \tag{A.10}$$

As with a Fuzzy system competition is engaged at the F_2 layer by way of a choice function that selects the F_2 unit J that has maximal $\mathbf{T}$ activity

$$T_J = max\{T_j : j = 1, \ldots, N\} \tag{A.11}$$

Top-down Hypothesis Testing

By choosing unit J an ART system makes a hypothesis about which category I belongs to. In a Gaussian system the matching function G_j (see eq. A.10) measures how close the current input vector is to category J. If this measurement is small enough in relation the systems vigilance measure; that is, if

$$G_j \geq \rho \tag{A.12}$$

then I is deemed to be close enough (a match) and learning is triggered (see below), however, if

$$G_j < \rho \tag{A.13}$$

then J is a bad hypothesis. In this case T_J is set to -1 (reset) in order that unit J can no longer compete to code I. A Search process is initiated whereby a new J is chosen, by equation A.11. The search continues until a J is chosen where equation A.12 is satisfied.

Learning

Once search ends, the count parameter n_J, and then sequentially the weight vectors μ_J and σ_J are updated thus

$$n_J = n_J + 1 \tag{A.14}$$

$$\mu_J = (1 - \frac{1}{n_J})\mu_J + \frac{I}{n_J} \tag{A.15}$$

$$\sigma_J = \sqrt{(1 - \frac{1}{n_J})\sigma_J^2 + \frac{(I - \mu_J)^2}{n_J}} \tag{A.16}$$

A.3 Fuzzy and Gaussian ARTMAP

A.3.1 Component Specification

An ARTMAP architecture incorporates two functionally equivalent unsupervised ART CMs (ART modules), ART$_a$, and ART$_b$, that are connected by a map field F^{ab}. A fuzzy system employs fuzzy ART modules, and likewise a Gaussian system uses Gaussian modules. The input **a** to ART$_a$ is a vector of size M^a $(\{a_1, \ldots, a_{M^a}\})$. The input **b** to ART$_b$ is a vector of size M^b $(\{a_1, \ldots, a_{M^b}\})$. Each element of **a** and **b** has a value in the interval [0,1].

The map field contains the same number of units N^b as the ART$_b$ F_2 field, and each F^{ab} unit k is linked by a single non-adaptive connection to a single F_2^b level unit k (where $k = 1, \ldots, N^b$). Its connection to the ART$_a$ module is facilitated by way of more adaptive LTM weights: each ART$_a$ F_2 level unit j has a weight vector w_j^* that connects it to all the units in the map field.

A.3.2 Model Parameters

Each ART module is controlled by way of its own set of intrinsic parameters (see sections A.1.2 and A.2.2 above), however, unlike in a stand-alone ART CM the ART$_a$ vigilance parameter ρ_a becomes an adaptive internally controlled component. At the beginning of each input presentation its value is set in accordance with a *baseline* vigilance parameter $\overline{\rho_a}$, which itself is set in the interval [0, 1] – employing the minimum value results in maximal ART$_a$ generalization. The map field also has its own vigilance parameter ρ_{ab} which is also set in the interval [0,1].

A further control parameter that is unique to fuzzy ARTMAP CMs is the match tracking parameter ε which is set in the range

$$-0.01 \leq \varepsilon \leq +0.01$$

where a small negative value improves generalization, but at a slight cost to accuracy, and vice versa for a small positive value. (see (Carpenter et al. 1996) for further detail).

A.3.3 Processing functions

Unsupervised Processing

The unsupervised modules ART$_a$ and ART$_b$ independently process the input vectors I^a and I^b respectively as per above with one exception, viz. a resonant state does not immediately initiate learning. The initiation of learning is ceded to map field control mechanisms.

Map Field Activation

The map field is activated whenever one of, or both of, the ART modules that in connects are in a state of resonance. The vector of activity z ($\{z_1, \ldots, z_{N^b}\}$) accross the map field is dependant upon the input it receives as follows:

- **Case 1:** If both ART modules are in the resonant state with the F_2^a unit J, and F_2^b unit K concurrently active, then

  $$z = y^b \wedge w_J^*$$

  where y^b is the F_2^b output vector.

- **Case 2:** If only ART$_a$ is in the resonant state with the F_2^a unit J solely active, then

  $$z = w_J^*$$

- **Case 3:** If only ART$_b$ is in the resonant state with the F_2^b unit K solely active, then

  $$z = y^b$$

Resonance or Mismatch

In the case where

$$|z| \equiv \sum_{k=1}^{N^b} z_k \geq \rho_{ab} \tag{A.17}$$

the full ARTMAP system is said to be in a resonant state. This state facilitates both prediction and learning (see below). Alternatively, if

$$|z| \equiv \sum_{k=1}^{N^b} z_k < \rho_{ab} \tag{A.18}$$

then there has been a predictive mismatch between the units J and K. The match tracking process (see below) is employed to remedy this problem.

Prediction and Learning

The course of prediction and learning is dependant not only on the satisfaction of equation A.17, but also upon the nature of the input that the map field is receiving:

- **Case 1:** If ART$_a$ and ART$_b$ are both active, then map field learning ensues as follows

  $$w_J^{*(new)} = \beta(y^b \wedge w_J^{*(old)}) + (1-\beta)w_J^{*(old)} \tag{A.19}$$

  In addition the weight values in each module are updated according to the appropriate equations (eq. A.6, or eqs. A.14-A.16). Finally, if match tracking has caused an increase in ART$_a$ vigilance its original value is reinstated

  $$\rho_a = \overline{\rho_a}$$

- **Case 2:** If only ART$_a$ is active, and prior map field learning has taken place then an F_2^a category J may predict an F_2^b category K, where K is the index of the F_2^b unit that recieves maximum top-down input (i.e. F^{ab} activity (z))

  $$T_K = max\{T_k : k = 1, \ldots, N^b\} \tag{A.20}$$

  In addition learning is initiated in the ART$_a$ module (eq. A.6, or eqs. A.14-A.16).

- **Case 3:** If only ART$_b$ is active, then case 2 applies but with the roles of each modules reversed.

Match Tracking and ART$_a$ Search

A predictive mismatch occurs in an ARTMAP system when an F_2^a unit J has already been active in conjuction with (and thus associated to by way of map field learning) an F_2^b unit k, where k does not have the same index as the F_2^b unit that is currently active (i.e. $k \neq K$). Given that the ART$_a$ input I^a should predict the ART$_b$ category K, the input cannot belong to ART$_a$ category J. Hence, the ARTMAP system resets unit J and initiates search for a more appropriate F_2^a category unit. This process is actioned by raising ART$_a$ vigilance just above the level of match between the current ART$_a$ input vector and the chosen F_2^a category representation as follows

In a fuzzy-based system:

$$\rho_a = \frac{|I^a \wedge w_J|}{|I^a|} + \varepsilon \tag{A.21}$$

In a Gaussian-based system:

$$\rho_a = \frac{n_J}{\Pi_{i=1}^{M^a} \sigma_{Ji}} \exp\left(-\frac{1}{2}\sum_{i=1}^{M^a}\left(\frac{I_i^a - \mu_{Ji}}{\sigma_{Ji}}\right)^2\right) \tag{A.22}$$

In either case, the ART$_a$ choice function (see equations A.4 and A.12) is no longer satisfied, which results in the ART$_a$ system inhibiting unit J and initiating a search process for a better category unit.

An example processing cycle for a fuzzy ARTMAP system is illustrated graphically in figure A.2.

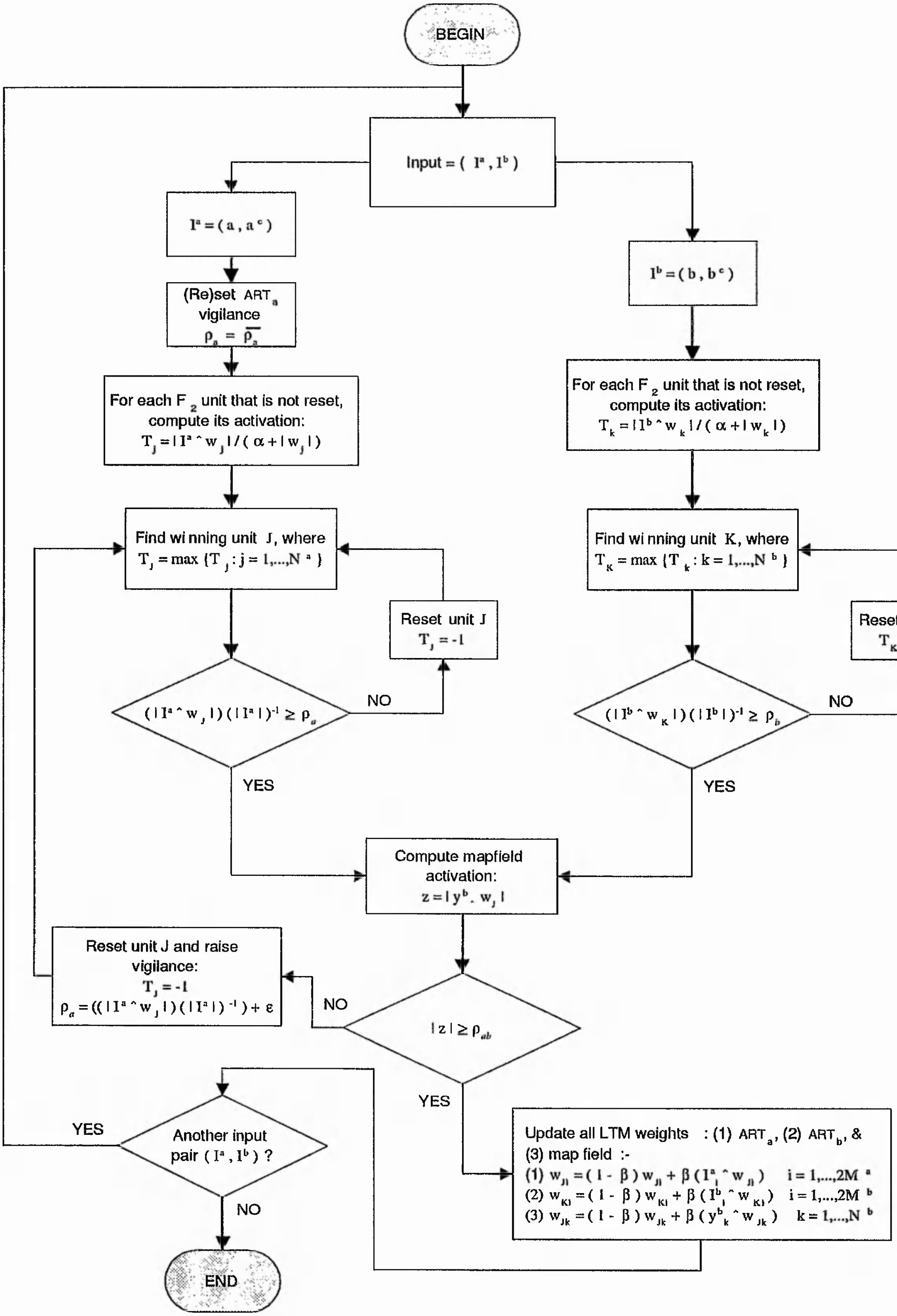
BEGIN
Input = (I^a , I^b)
I^a = (a , a^c)
(Re)set ART_a vigilance
$\rho_a = \overline{\rho_a}$
For each F_2 unit that is not reset, compute its activation:
$T_j = |I^a \wedge w_j| / (\alpha + |w_j|)$
Find winning unit J, where
$T_J = \max\{T_j : j = 1,\ldots,N^a\}$
Reset unit J
$T_J = -1$
$(|I^a \wedge w_J|)(|I^a|)^{-1} \geq \rho_a$
NO
YES
I^b = (b , b^c)
For each F_2 unit that is not reset, compute its activation:
$T_k = |I^b \wedge w_k| / (\alpha + |w_k|)$
Find winning unit K, where
$T_K = \max\{T_k : k = 1,\ldots,N^b\}$
Reset unit K
$T_K = -1$
$(|I^b \wedge w_K|)(|I^b|)^{-1} \geq \rho_b$
NO
YES
Compute mapfield activation:
$z = |y^b . w_J|$
Reset unit J and raise vigilance:
$T_J = -1$
$\rho_a = ((|I^a \wedge w_J|)(|I^a|)^{-1}) + \varepsilon$
NO
$|z| \geq \rho_{ab}$
YES
Update all LTM weights : (1) ART_a, (2) ART_b, & (3) map field :-
(1) $w_{Ji} = (1 - \beta) w_{Ji} + \beta (I^a_i \wedge w_{Ji})$ $i = 1,\ldots,2M^a$
(2) $w_{Ki} = (1 - \beta) w_{Ki} + \beta (I^b_i \wedge w_{Ki})$ $i = 1,\ldots,2M^b$
(3) $w_{Jk} = (1 - \beta) w_{Jk} + \beta (y^b_k \wedge w_{Jk})$ $k = 1,\ldots,N^b$
YES
Another input pair (I^a , I^b) ?
NO
END

Figure A.2: Shows the learning process in a fuzzy ARTMAP model.

Appendix B

Chapter Six – Further Data

This appendix contains further data that supports the information presented in chapter 6.

B.1 Full GA Sales Results

The following section contains the full results of the investigations described in section 6.1.4. In the main body of this thesis these results are summised in table 6.7.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	210 (94.59%)	222 (100.0%)	222 (100.0%)	210 (94.59%)
GA 2	207 (93.24%)	222 (100.0%)	222 (100.0%)	207 (93.24%)
GA 3	209 (94.14%)	222 (100.0%)	222 (100.0%)	209 (94.14%)
GA 4	208 (93.69%)	222 (100.0%)	222 (100.0%)	208 (93.69%)
GA 5	213 (95.95%)	222 (100.0%)	222 (100.0%)	213 (95.95%)
Average	209.4 (94.32%)	222 (100.0%)	222 (100.0%)	209.4 (94.32%)

Table B.1: 12% Noisy Test Data.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	202 (90.99%)	222 (100.0%)	222 (100.0%)	202 (90.99%)
GA 2	198 (89.19%)	222 (100.0%)	222 (100.0%)	198 (89.19%)
GA 3	210 (94.59%)	222 (100.0%)	222 (100.0%)	210 (94.59%)
GA 4	207 (93.24%)	222 (100.0%)	222 (100.0%)	207 (93.24%)
GA 5	206 (92.79%)	222 (100.0%)	222 (100.0%)	206 (92.79%)
Average	204.6 (92.16%)	222 (100.0%)	222 (100.0%)	204.6 (92.16%)

Table B.2: 14% Noisy Test Data.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	210 (94.59%)	222 (100.0%)	222 (100.0%)	210 (94.59%)
GA 2	213 (95.95%)	221 (99.55%)	221 (99.55%)	213 (95.95%)
GA 3	200 (90.09%)	222 (100.0%)	222 (100.0%)	200 (90.09%)
GA 4	199 (89.64%)	222 (100.0%)	222 (100.0%)	199 (89.64%)
GA 5	206 (92.79%)	222 (100.0%)	222 (100.0%)	206 (92.79%)
Average	205.6 (92.61%)	221.8 (99.91%)	221.8 (99.91%)	205.6 (92.61%)

Table B.3: 16% Noisy Test Data.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	190 (85.59%)	222 (100.0%)	222 (100.0%)	190 (85.59%)
GA 2	191 (86.04%)	221 (99.55%)	221 (99.55%)	191 (86.04%)
GA 3	193 (86.94%)	222 (100.0%)	222 (100.0%)	193 (86.94%)
GA 4	191 (86.04%)	222 (100.0%)	222 (100.0%)	191 (86.04%)
GA 5	182 (81.98%)	221 (99.55%)	221 (99.55%)	182 (81.98%)
Average	189.4 (85.32%)	221.6 (99.82%)	221.6 (99.82%)	189.4 (85.32%)

Table B.4: 18% Noisy Test Data.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	187 (84.23%)	222 (100.0%)	222 (100.0%)	187 (84.23%)
GA 2	179 (80.63%)	222 (100.0%)	222 (100.0%)	179 (80.63%)
GA 3	188 (84.68%)	221 (99.55%)	221 (99.55%)	188 (84.68%)
GA 4	190 (85.59%)	220 (99.10%)	220 (99.10%)	190 (85.59%)
GA 5	188 (84.68%)	221 (99.55%)	221 (99.55%)	188 (84.68%)
Average	186.4 (83.96%)	221.2 (99.64%)	221.2 (99.64%)	186.4 (83.96%)

Table B.5: 20% Noisy Test Data.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	180 (81.08%)	221 (99.55%)	221 (99.55%)	180 (81.08%)
GA 2	181 (81.53%)	221 (99.55%)	221 (99.55%)	181 (81.53%)
GA 3	180 (81.08%)	222 (100.00%)	222 (100.00%)	180 (81.08%)
GA 4	173 (77.93%)	220 (99.10%)	220 (99.10%)	173 (77.93%)
GA 5	183 (82.43%)	221 (99.55%)	221 (99.55%)	183 (82.43%)
Average	179.4 (80.81%)	221.0 (99.55%)	221.0 (99.55%)	179.4 (80.81%)

Table B.6: 22% Noisy Test Data.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	169 (76.13%)	217 (97.75%)	217 (97.75%)	169 (76.13%)
GA 2	179 (80.63%)	220 (99.10%)	220 (99.10%)	179 (80.63%)
GA 3	176 (79.28%)	222 (100.00%)	222 (100.00%)	176 (79.28%)
GA 4	175 (78.83%)	221 (99.55%)	221 (99.55%)	175 (78.83%)
GA 5	169 (76.13%)	220 (99.10%)	220 (99.10%)	169 (76.13%)
Average	173.6 (78.20%)	220.0 (99.10%)	220.0 (99.10%)	173.6 (78.20%)

Table B.7: 24% Noisy Test Data.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	164 (73.87%)	221 (99.55%)	221 (99.55%)	164 (73.87%)
GA 2	154 (69.37%)	221 (99.55%)	221 (99.55%)	154 (69.37%)
GA 3	152 (68.47%)	220 (99.10%)	220 (99.10%)	152 (68.47%)
GA 4	162 (72.97%)	218 (98.20%)	218 (98.20%)	162 (72.97%)
GA 5	146 (65.77%)	220 (99.10%)	220 (99.10%)	146 (65.77%)
Average	155.6 (70.10%)	220.0 (99.10%)	220.0 (99.10%)	155.6 (70.10%)

Table B.8: 26% Noisy Test Data.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	145 (65.32%)	217 (97.75%)	217 (97.75%)	145 (65.32%)
GA 2	140 (63.06%)	217 (97.75%)	217 (97.75%)	140 (63.06%)
GA 3	140 (63.06%)	221 (99.55%)	221 (99.55%)	140 (63.06%)
GA 4	138 (62.16%)	217 (97.75%)	217 (97.75%)	138 (62.16%)
GA 5	144 (64.86%)	220 (99.10%)	220 (99.10%)	144 (64.86%)
Average	141.2 (63.60%)	218.4 (98.38%)	218.4 (98.38%)	141.2 (63.60%)

Table B.9: 28% Noisy Test Data.

	Recall of VISUAL Knowledge		Recall of LEXICAL Knowledge	
	Lexical Input (Identification)	Visual Input	Visual Input (Description)	Lexical Input
GA 1	123 (55.41%)	213 (95.95%)	213 (95.95%)	123 (55.41%)
GA 2	131 (59.01%)	213 (95.95%)	213 (95.95%)	131 (59.01%)
GA 3	125 (56.31%)	214 (96.40%)	214 (96.40%)	125 (56.31%)
GA 4	142 (63.96%)	214 (96.40%)	214 (96.40%)	142 (63.96%)
GA 5	121 (54.50%)	216 (97.30%)	216 (97.30%)	121 (54.50%)
Average	128.4 (57.84%)	214.0 (96.40%)	214.0 (96.40%)	128.4 (57.84%)

Table B.10: 30% Noisy Test Data.

	Training Times		Operating Times	
	total	per pattern	total	per pattern
1	2.58	0.070	4.07	0.018
2	2.75	0.074	4.12	0.019
3	2.64	0.071	3.95	0.018
4	2.53	0.068	3.95	0.018
5	2.52	0.068	4.12	0.019
6	2.69	0.073	4.12	0.019
7	2.74	0.074	4.18	0.019
8	2.69	0.073	4.12	0.019
9	2.64	0.071	4.06	0.018
10	2.58	0.070	4.17	0.019
11	2.69	0.073	4.23	0.019
12	2.64	0.071	4.12	0.019
13	2.53	0.068	4.12	0.019
14	2.58	0.070	4.39	0.020
15	2.69	0.073	4.12	0.019
16	2.59	0.070	4.17	0.019
17	2.52	0.068	4.12	0.019
18	2.58	0.070	4.18	0.019
19	2.58	0.070	4.34	0.020
20	2.58	0.070	4.01	0.018
21	2.58	0.070	4.12	0.019
22	2.53	0.068	4.12	0.019
23	2.74	0.074	4.34	0.020
24	2.64	0.071	4.17	0.019
25	2.63	0.071	4.18	0.019

Table B.11: Times (in seconds) taken for ARTMAP models (1-25) to learn and recall training patterns.

	Training Times		Operating Times	
	total	per pattern	total	per pattern
26	2.63	0.071	4.12	0.019
27	2.86	0.077	4.17	0.019
28	2.52	0.068	4.18	0.019
29	2.53	0.068	4.12	0.019
30	2.69	0.073	4.23	0.019
31	2.64	0.071	4.17	0.019
32	2.53	0.068	4.28	0.019
33	2.69	0.073	4.23	0.019
34	2.69	0.073	4.12	0.019
35	2.53	0.068	4.17	0.019
36	2.69	0.073	4.23	0.019
37	2.59	0.070	4.22	0.019
38	2.58	0.070	4.12	0.019
39	2.64	0.071	4.12	0.019
40	2.59	0.070	4.28	0.019
41	2.63	0.071	4.40	0.020
42	2.63	0.071	4.34	0.020
43	2.58	0.070	4.23	0.019
44	2.58	0.070	4.18	0.019
45	2.58	0.070	4.17	0.019
46	2.53	0.068	4.34	0.020
47	2.58	0.070	4.67	0.021
48	2.64	0.071	4.45	0.020
49	2.58	0.070	4.45	0.020
50	2.63	0.071	4.84	0.022

Table B.12: Times (in seconds) taken for ARTMAP models (26-50) to learn and recall training patterns.

Appendix C

Publications

The following papers have been published based upon the work detailed in this thesis:

1. Chandler, N., Balendran, V., Evett, L., and Sivayoganathan, K. (1995). From Speech Recognition to Language Understanding: An Engineering Perspective. *11th National Conference on Manufacturing Research*, 562-566. DeMontfort University, England.

2. Chandler, N., Balendran, V., Evett, L., and Sivayoganathan, K. (1996). From Hearing to Understanding. *Manufacturing Engineer*, **75**(2), 65-68.

3. Chandler, N., Balendran, V., Evett, L., and Sivayoganathan, K. (1997). On the Weaknesses of Arbitrary Connectionist Representations. *Forth Neural Computation and Psychology Workshop; Connectionist Representations : Theory and Practice.* University of London, England.

4. Chandler, N., Balendran, V., Evett, L., and Sivayoganathan, K. (1997). On the Importance of Reality in Representations. In Riegler, A., and Peschl, M. (eds.), *Does Representation need Reality? - Proceedings of the International Conference 'New Trends in Cognitive Science'.* Austrian Society for Cognitive Science Technical Report 97-01, Vienna, Austria.

5. Chandler, N., Balendran, V., Evett, L., and Sivayoganathan, K. (1999). Reality: A Prerequisite to Meaningful Representation. In Riegler, A., and Peschl, M. (eds.), *Understanding Representation in the Cognitive Sciences: Does Representation Need Reality?.* NY: Kluwer Academic/Plenum Publishers.