Cascade-Correlation as a Model of Representational Redescription

J. K. Brook

CSRP 491

July, 1998

ISSN 1350–3162

Cognitive Science
Research Papers

Acknowledgements

I would like to thank Chris Thornton for his supervision, and the members of my review committee, Ben du Boulay, Maggie Boden and Ron Chrisley for useful comments on my early work. Thanks are also due to Scott Fahlmann and Mike Mozer for providing useful information on their work, and to Steven Philips and Angelo Cangelosi. Julian Budd, Jason Noble and Nick Ross provided many useful comments on drafts of the thesis. I would also like to thank Matthew Hennessy, Linda Thompson and the systems staff at COGS for all their help over the years. Thanks also to my examiners Andy Clark and Ron Chrisley for their useful feedback on this thesis.

I would also like to thank all who have provided a mixture of support, friendship and distraction during my time in COGS, and during the completion of my write-up in Edinburgh, in particular Stephen Eglen, Alan Jeffrey, Richard Dallaway, Philip Jones, Stephen Dunn, Jason Noble, Susi Ross, Adrian Thompson, Clive Cox, Julian Budd, Marisa Ueda and Eevi Beck. Thanks also to Dave Berry and all at Harlequin Ltd. in Edinburgh for bearing with me while I finished this thesis.

I am indebted to my grandparents, Julian and Joan Smith, as well as my parents for providing me with the means to undertake this project, and also to COGS for supporting me by means of a fee waiver.

Cascade-Correlation as a Model of Representational Redescription

J. K. Brook

Abstract

How does knowledge come to be manipulable and flexible, and transferable to other tasks? These are issues which remain largely untackled in connectionist cognitive modelling.

The Representational Redescription Hypothesis (RRH) (Karmiloff-Smith, 1992b) presents a framework for the emergence of abstract, higher-order knowledge, based on empirical work from developmental psychology. The RRH claims that during learning/development initially-implicit knowledge is rendered progressively more explicit via the reiterated action of the re-description process, resulting in a hierarchy of increasingly explicit and accessible representations.

This thesis focuses on investigating in practice claims made for connectionism as a model of redescription (e.g., Clark and Karmiloff-Smith (1993)) and on applying methods from recent work in developmental connectionism to the construction of a computational model of RR. The modelling effort centres on a constructive incremental architecture — cascade-correlation (CC) (Fahlman & Lebiere, 1990) — which produces a conservative hierarchy of increasingly high-level representations as the RRH proposes.

Two main models are presented. The first is designed to capture a feature of children's comprehension of the French article system (Karmiloff-Smith, 1979a). Redescriptive effects are seen here in the changing functional status of article representations as well as in symptomatic behavioural errors. Resource-phasing was also applied to two important internal parameters of CC.

The second model aims to capture the effects of redescription on sequence learning. Recurrent CC was trained to count, and to give and compare the cardinalities of small series of stimuli. Accessibility of representations was assessed here through task transfer. Despite some success in capturing transfer, a short complementary study of structural transfer between networks learning formal grammars suggested that positive transfer in CC depends on perceptual similarity as in other supervised connectionist schemes.

The models also address constraints on RR such as the timing and triggering of redescription and the ordering of representational formats.

A brief comparative study of *skeletonisation* is also presented.

Submitted for the degree of D. Phil.

University of Sussex

April, 1997

Contents

Chapter 1

Introduction

What is abstract thought? How are we able to manipulate concepts, to reflect upon them, and redeploy them in novel contexts? How does what we have learned about a particular subject become integrated and systematic? It is these questions which provide the motivation for this project.

The project takes a dynamic and developmental approach to constructing computational models of the mechanisms underlying flexible and systematic thought, attempting to model how these might emerge from and interact with cognitive change.

The choice of a developmental framework in which to consider these questions reflects a general movement (relatively recent in cognitive science (Clark, 1993a, p. ix)) away from the consideration of mature human adult competence in isolation, and towards an approach which is broader in explanatory scope, and in which dynamics plays a central role. An important consequence of this approach is that development has a role, not merely as a domain to be modelled, but also as a means of studying knowledge and cognition.

The modelling effort is guided by a hypothesis — the Representational Redescription Hypothesis (Karmiloff-Smith, 1992b) — whose focus is the emergence of abstract concepts from procedural knowledge over the course of both learning and development in humans.

The models presented have all been constructed within a connectionist framework[1]. The link between connectionist modelling and the RRH was first made by Clark and Karmiloff-Smith (1993) and Karmiloff-Smith (1992b) and can be seen as working in two directions. Firstly, despite a lack of detailed commitment in the RRH itself concerning concrete mechanisms underlying cognitive change, connectionism has been observed by Karmiloff-Smith to exhibit certain important correspondences with the RR model (discussed in more detail in section 1.1 below) in terms of the style in which task-knowledge is represented and the operations facilitated by that representation. Secondly, the RRH seems to encompass and reiterate requirements for a model of human cognition which have traditionally been considered problematic for connectionist modelling, in particular the systematic reuse of concepts for learning structurally related tasks.

This project is aimed at examining the claims made for connectionism as a model of redescription and focuses on machine-learning techniques known collectively as *resource phasing* or *incremental learning*, which have been successfully used to model other aspects of cognitive development.

The remainder of this chapter presents an overview of the RRH, as well as discussing the claims for connectionism as a model of the mechanisms which might underlie it. The specific connectionist architectures and incremental methods used are then presented and motivated as

[1]Familiarity with connectionism is assumed throughout this thesis. See Plunkett and Sinha (1992) for a brief introduction tailored to developmental models

well as the example task domains on which the modelling focuses. The chapter concludes with a summary of the contributions made by this thesis.

1.1 The Representational Redescription Hypothesis

The Representational Redescription Hypothesis (Karmiloff-Smith, 1992b) is an attempt to account for certain qualitative phenomena observed in development and in child- and adult learning, in particular the progression from knowing how to do something (procedural knowledge) to being able to reflect upon that knowledge, discuss it and manipulate it. What is important is not that performance at a task actually changes but rather that the status of the knowledge to the learner changes, i.e., the form in which it is understood and its integration into the rest of the learner's knowledge. Indeed overall task performance may actually worsen symptomatically as the roles of knowledge change and conflict.

In attempting to explain this progression, the hypothesis puts forward a series of *phases* during each of which knowledge is thought of as being represented and stored in different formats (see figure 2.1). Each phase is more explicit than the last (at the highest level this is linked with verbalisability) and allows progressively more flexibility in its use, whether verbal or non-verbal, in particular in facilitating further learning on other tasks or domains.

The three phases are known as **I** (implicit) at which knowledge is procedurally represented and unavailable outside the original input–output mapping, **E1**, an intermediate level of the so-called 'explicitation' process, at which knowledge is more accessible than at the first phase, but still not verbalisable, and **E2/3** conscious (and possibly verbalisable) explicit knowledge. Transition between these formats is hypothesised to involve the reiterative *redescription* of previous representations, which while not specified in detail, seems to 'reduce' the knowledge by discarding some of the original detail. It is also hypothesised that the generation of new formats is conservative or redundant — rather than each new format supplanting the last, representations form a hierarchy of levels at which the same knowledge is differently represented. An important aspect of the hypothesis is the emphasis it places on endogenous (or internally driven) change; although learning happens with respect to external influences, representations are assumed to change 'off-line' after (at least some) initial competence at the task has been achieved, rather than in response to external pressures, such as the need to improve task performance. Rather than being provoked to improve by the presence of instability, the system is driven to reappropriate already stable states.

An informal example of this progression presented by Karmiloff-Smith (1992b) is that of learning to play a musical piece. First one must learn to reach the initial mastery of having assembled notes and phrases into a continuous piece. Playing at this point is competent but relatively inflexible — maybe the volume of the whole piece may be adjusted but little else. The RRH has it that with time, redescription acts to increase the flexibility and accessibility of the knowledge and its components. This manifests itself in effects such as the ability to begin playing at arbitrary points during a piece, to add subtle emphases, and to improvise upon it.

Chapter 2 describes the RRH in more detail, presenting supporting empirical work done by Karmiloff-Smith, and surveying it in the context of other developmental hypotheses and direct responses.

1.2 Connectionism, Developmental Modelling and the RRH

The RRH is put forward as 'a framework — rather than a precise theory — for exploring possible generalities in developmental change across a range of domains.' (Karmiloff-Smith, 1994), and makes no detailed commitment to any possible mechanisms for redescription itself. However, in Clark and Karmiloff-Smith (1993) a set of general requirements are put forward: any model should spontaneously come to manipulate its own representations, preserving the results of previous learning, and should form new structured representations of its knowledge which can

be manipulated by other processes. In the light of these requirements Karmiloff-Smith (1992b), Clark and Karmiloff-Smith (1993) models based on connectionism were put forward as the most promising of the computational modelling paradigms then current.

The link between the hypothesis and connectionism was prompted in part by the observation (Karmiloff-Smith, 1992b) of an apparent similarity between the opaqueness and embeddedness of procedural knowledge at the initial level of the RR model and the kind of knowledge captured by a trained network. But as Clark and Karmiloff-Smith (1993) point out, standard error-driven networks do not spontaneously go beyond this success to further link, systematise and redeploy their knowledge as the RRH demands.

In response to this point, several (existing) schemes for augmenting standard networks have been proposed in the literature as models for RR. These suggestions are discussed at greater length in Chapter 3, but briefly these have generally involved structural manipulations on networks, such as copying whole or partial networks during or after training and using these and other means to capture the idea of re-representing and redeploying knowledge.

Techniques such as these, which involve the qualitative control of learning by manipulating the structure of the network or its training data are examples of *incremental-learning* or *resource-phasing* methods. What unites such methods in general is the manipulation over the course of evolution, development or learning of the data seen by the agent.

Resource-phasing techniques have already been explored as part of work on connectionist developmental modelling (e.g., Elman (1991), Plunkett and Marchman (1991)). Recent years have seen a growth in interest in the use of connectionist models to capture developmental change (McClelland (1989), Plunkett and Sinha (1992), Bates and Elman (1992)). A central motivation for this is the ability of connectionist systems (in contrast with most symbolic systems) to capture change, especially qualitative change, emergently within a unitary framework. This provides us with a means of modelling process and thus with a concrete testbed for theories of cognitive change (Simon & Halford, 1995). Chapter 3 focuses on developmental connectionism as well as covering relevant issues from connectionist efforts to model cognition in general, in particular representation, systematicity, and explicitness.

The experimental work reported in this thesis explores the applicability of existing connectionist developmental modelling techniques to the construction of a computational model of the RRH, focusing on the claim implicit in Clark (1993a) that resource phasing in networks supports the progressive explicitation and redescription of the products of learning that the RRH requires. These efforts are focused in turn on a particular resource phasing scheme — the Cascade-correlation architecture (Fahlman & Lebiere, 1990).

Cascade correlation is an example of a constructive (or generative) architecture — its policy is to 'recruit' hidden units over the course of learning. These are installed hierarchically into the network. Shifts in on-line processing and representational power are controlled in a gradualistic, incremental way which has made it suitable for developmental modelling (see Shultz, Schmidt, Buckingham, and Mareschal (1995) for an overview of this work). Cascade-correlation is particularly promising as a model of redescription as it is not only incremental, but hierarchical and conservative (the results of previous learning are preserved and mediate subsequent learning). Another obvious analogue between it and the RR model is that it also uses separate mechanisms for on- and off-line training (based on reducing error and maximising correlation respectively) integrating these within a single framework. Cascade-correlation is described in more detail in Chapter 4.

A short complementary study of the selectionist resource-phasing scheme *skeletonisation* (Mozer & Smolensky, 1989a, 1989b) is presented in chapter 7 in the context of the article-function task. Skeletonisation prunes units from previously trained networks according to a measure of relevance not directly based on error-reduction.

1.3 Example domains

The experimental work reported in this thesis is based around two tasks drawn from those investigated by Karmiloff-Smith in the domains of language and number. The first is based on the learning of the correspondences between articles and their functions presented in Karmiloff-Smith (1979a). The second looks at counting and cardinality as an example of sequence-learning domains.

The models of the two tasks also complement each other in exemplifying differing domain-general constraints, in terms of their use of recurrent or non-recurrent architecture, as well as in the domain-specific constraints inherent in the design and biasing of training examples, discussed for each model. Beyond these differences, the models are constructed within the same learning framework, allowing both comparison and discussion of the possibility of constructing a domain-general connectionist model of RR.

1.3.1 Redescriptive effects in the acquisition of the French article system

Children's usage and comprehension of the definite and indefinite articles in French exhibits a U-shaped behavioural curve (Karmiloff-Smith, 1979a) as they try to reconcile the non-specific ('a') and specific ('one') functions of the indefinite article ('un'/'une'). Karmiloff-Smith (1992b) explains this within the framework of the RR model. Initially the articles are represented individually, obeying a one-form–one-function constraint, but as redescription renders the commonalities between these representations explicit, overmarking appears in production as well as errors in comprehension. Eventually these two representations are reconciled, giving a *pluri-functional* status to articles and resulting again in correct behaviour.

Cascade-correlation was used to model these redescriptive phenomena on comprehension. The effects of training-set bias on learning trajectory was examined, as well as the effects of different proportions of externally and internally driven learning. This work is presented in Chapter 5.

1.3.2 Sequence learning and the RRH

The second set of experimental work presented focuses on temporal sequence learning, in particular in the domain of counting. This class of tasks was chosen for a number of reasons. Firstly the redescriptive phenomena associated with it are observed in a cross-section of domains studied by Karmiloff-Smith and it is thus useful in examining whether RR effects are underlain by a unitary mechanism. Sequence-learning is also associated with a well-defined set of effects in the RRH literature — for instance it is claimed that redescription first makes end-most components of sequences accessible before individuating inner components. In the counting domain this is claimed to explain how the concept of number gained through counting is linked to that of cardinality. Recurrent cascade-correlation was trained to count short sequences of stimuli, to output the cardinality of a sequence without producing the intermediate counting outputs, and to compare the cardinalities of two consecutive sequences. Transfer between each of these tasks was also investigated.

To control for the role of perceptual cues in the results of the transfer experiments between counting cardinality and comparison, RCC was also trained on strings generated by regular grammars, i.e., to approximate the behaviour of deterministic finite-state automata. Unlike the counting task correct performance on this task requires the network to attend to the identities of individual stimuli during learning and to abstract from these when trained to induce a machine which is isomorphic but differently labelled. Another motivation for this study was comparison with previous work on learning and structural transfer using discrete locally recurrent networks (e.g., Cleeremans (1993), Chrisley (1993), Dienes, Altmann, and Gao (1995), Jackson and Sharkey (1995)). Chapter 6 presents the sequence-learning studies in more detail.

In terms of formats, the aim in all the models reported here has been to capture the progression from level I to level E1 — the modelling of accessibility to consciousness or verbal expression was considered to be outside the scope of this project. The models are also designed to capture the overall dynamics of the behavioural progressions in each domain.

1.4 Contributions of this thesis

This thesis presents the first study dedicated to investigating the claims that connectionist architectures can provide models for the RRH in the context of particular domains discussed as evidence for RR effects by Karmiloff-Smith, specifically sequence-learning (exemplified by counting) and language acquisition. In particular it investigates whether a class of such architectures — those which are both incremental and error-driven — are particularly suited to this modelling effort. It is also the first practical investigation of network transfer as the operationalisation of the progressive accessibility characteristics of the RRH.

The playroom experiment extends the range of incremental learning techniques which have been used in developmental models based on cascade-correlation. Specifically, the patience and candidate pool-size parameters were varied over the course of training in an attempt to control the timing and nature of qualitative representational and behavioural change as well as to capture the early one-form–one-function constraint.

The study of counting, cardinality and comparisons was the first use of recurrent cascade-correlation in constructing a developmental model of temporal behaviour. The application of cascade-correlation to structural transfer between isomorphic but re-labelled finite-state machines was also novel.

The short study using skeletonisation was the first application of this technique in an attempt to model the RRH. The augmentation of the technique with weight freezing and network copying was a novel extension to skeletonisation.

Chapter 2

The Representational Redescription Hypothesis

2.1 Introduction: the Representational Redescriptional Hypothesis

The Representational Redescription Hypothesis (RRH) (Karmiloff-Smith, 1986, 1992b) is a set of related claims about qualitative behavioural change during development, child learning and also adult learning in some cases. It is concerned with the progression from competent performance of a skill (simply, knowing how to perform a task, such as balancing objects on a fulcrum or producing mature usage of personal pronouns), to the ability to reflect upon, discuss and manipulate that knowledge.

2.1.1 Implicit and explicit representations

Representations, in the terms of the RRH, are considered to be that which sustains behaviour in a particular domain. The RRH also proposes that the learning of a particular task can be divided into a number of serially ordered *phases*. It is assumed that the initial phase of learning, which results in (at least partially) successful performance at a task (or *behavioural mastery*), is associated with an implicit procedural representational format. The hypothesis states that the final phase of learning, at which knowledge is accessible to introspection and verbal expression, is underlain by an explicit representational format. If these two levels are viewed in isolation, the hypothesis is similar to many other theories of learning and mental representation (e.g., Mandler (1988)) in that it proposes that unconscious procedural behaviour involves implicit representation while conscious behaviour involves explicit representations. Where the RRH differs from these accounts, and what makes it of particular interest to connectionist modellers in my view, is that its main concern is with the actual process of transition between these two types of representational formats, and identifies explicitly intermediate states between them on the basis of experimental evidence.

Domains

Before continuing discussion of the general characteristics of the RRH, it is necessary to define the term *domain* as used in this context. A domain has two important senses here. The first is that it is to be contrasted with the Fodorean notion of a *module* (Fodor, 1983). Specifically the latter is an 'information-processing unit that encapsulates . . . knowledge and the computations on it' (Karmiloff-Smith, 1992b, p. 6), while a domain is a 'set of representations sustaining a specific area of knowledge', examples being language, number and physics. The process of RR is seen as acting domain-specifically but this does not imply that knowledge is modularised according to these domains.

The second reason for defining RR with respect to domains is that, unlike representational change in a stage theory, RR occurs at different times with respect to different areas of knowledge and thus some way of distinguishing between these is needed.

The granularity of domains may be partially defined with respect to the related notion of *micro-domains*. These are smaller groups of related tasks, skills, or bodies of knowledge, such as pronoun-acquisition, or basic knowledge of gravity, and are seen as being subsets of the domains (of language and physics respectively).

As Keil (1990) observes, the notion of domain varies considerably between different theorists. To some extent, the knowledge considered to form a domain relates to the constraints which apply to it. Despite variations in the breadth of knowledge circumscribed by a given set of constraints, the common factor is that 'domain specific constraints are predicated on specific sorts of knowledge types and do not blindly constrain any possible input to learning.' (Keil, 1990, p. 139), and Keil goes on to add that the working definition of a domain is 'in terms of patterns of learning' — if restrictions to possible solutions are unique to a particular body of knowledge, that knowledge is considered a domain.

2.1.2 Accessibility of knowledge

Closely related to the process of increasing explicitness is that of increasing accessibility — knowledge and representations at the initial level are 'bracketed' and unavailable outside the original input–output mapping, and become increasingly widely accessible over the course of the so-called *explicitation* process, central to RR. At the intermediate level these representations become available to other tasks within the same domain, and at the final level transferable between domains.

2.1.3 Sources of knowledge

Another of the fundamental assumptions from which the RRH proceeds is that the knowledge stored in the mind derives from several different sources. The first is environmental feedback, which allows us to learn from failure in achieving some action, while a second external source of information is provided by linguistic statements made by others. The focus of the RRH is on the way internal sources of knowledge, in particular the redescribed versions of mastered tasks, are pressed into service by learners.

These three factors — the transition from implicit to explicit, the increase in accessibility and the emphasis on internal sources of knowledge (and knowledge change) — are key aspects of the RRH. The next section presents the phases of the RR model in more detail, describing experimental evidence for the associated effects.

2.2 The RR Model

In attempting to explain the implicit–explicit progression, Karmiloff-Smith (1986, 1992b) has formulated the RR model (see figure 2.1). This proposes that over developmental time, knowledge about a task or domain comes to be represented as a hierarchy. In more recent formulations (e.g., Karmiloff-Smith (1992b)) this comprises at least four different formats known as **I** — implicit, **E1** — the first explicit level, **E2** — at which knowledge is explicit and conscious but non-verbalisable, and **E3** — at which it becomes explicit, conscious and verbalisable.

It is hypothesised that transition between these formats involves the reiterated action of the process of *representational redescription* from which the hypothesis takes its name. The phases mentioned above refer to periods between and including such transitions. A given learner may simultaneously be in different phases with respect to different tasks or domains, thus these phases are domain-specific and should be contrasted with Piagetian stages (Piaget, 1953). Each of these phases and the representational format associated with it is considered in more detail below.

The Implicit Phase

At the initial level representations are thought of as being implicit and orientated towards the task of 'responding to and analyzing stimuli in the external environment' (Karmiloff-Smith,

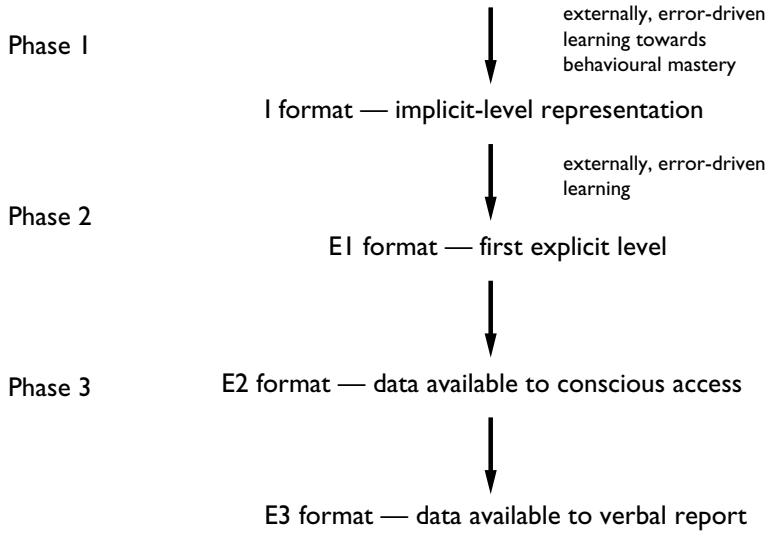


Figure 2.1: The Representational Redescription Model

1993). The hypothesis also has it that in this phase new knowledge about a task is added piecemeal through a process Karmiloff-Smith (1990, 1992b) refers to as 'representational adjunction'. The idea is that, once stable, representations corresponding to microdomains, tasks or parts of tasks accumulate independently of each other and without regard to possible redundancy, giving rise to characteristic I-level behaviour, which is efficient and at least partially successful in that it sustains competent performance at a task, but is neither flexible in its use, linked to or usable by representations of other similar tasks, or available for verbal expression. It is important to note that performance at this level may closely resemble eventual performance; it may only be the underlying representations which differ.

The Implicit Format

The hypothesis proposes that underlying such behaviour is an implicit representational format. In addition to being inaccessible beyond the task itself and stored independently of each other, representations in this format also conform to the constraints that information is encoded in procedural form, and that these procedures are 'sequentially specified' (Karmiloff-Smith, 1992b, p. 20).

Rutkowska (1993) distinguishes between the sense of the term 'procedural' in computing (where it is set in opposition to declarative, particularly in traditional AI) and as it is used in developmental psychology. In this context, knowledge which is action-based and perhaps available to other subsystems only implicitly in special-purpose procedures is considered to be procedurally represented (Rutkowska, 1993). Karmiloff-Smith (1990) adds to this definition constraints such as serial ordering, which are considered to be *domain-general* or to act independently of particular domains.

Such procedures are available to other operators (such as other procedures) only as wholes, and the hypothesis has it that it is only after redescription that their components become available (although the exact force of the term 'components' is not made precise outside the context of specific tasks).

Examples

The performance of children on the task of balancing a set of visibly or invisibly weighted blocks exhibits a 'U-shaped behavioural curve' (Strauss & Stavy, 1982). Karmiloff-Smith and Inhelder (1975) observed that the children of 4 and 8 years they studied were able to balance any blocks presented to them, while at intermediate ages some were rejected as 'unbalancable'.

In this microdomain, the initial phase provides concrete examples of two characteristics of

the implicit phase and format. Firstly the performance of the youngest children is data-driven — there is no apparent sensitivity to anything but the observable data, which here takes the form of proprioceptive and visual feedback. Secondly, this microdomain provides evidence for the idea of representational adjunction — information was seemingly not generalised from a block to other identical blocks; rather the balancing of each block was treated as an isolated problem.

In the microdomain of counting, we see again that at the implicit level children 'run off' a mastered procedure. Karmiloff-Smith (1992b) reports that young children who have mastered counting up to small numbers are like the youngest group in the block-balancing study, in that they go through the whole procedure again when faced with an identical display. These effects are symptomatic of the action of sequential representational constraints.

Also, despite knowing the relevance of the question 'How many?' to performing a count, these children do not recognise that there is a component of the count which corresponds to the cardinality of the set to be counted. In the terms of the hypothesis 'the knowledge embedded in the procedure is not yet available as separate components' (Karmiloff-Smith, 1992b, p. 104).

In the language domain the initial level is generally characterised in the studies carried out by Karmiloff-Smith (1979b, 1992b) by correct usage, based, as in the block-balancing microdomain, on superficial characteristics of the input data. For instance, although by the age of around 3, children are able both to perceive and produce words, their idea of what words *are* is comparatively impoverished, with closed-class words (e.g., 'the', 'of') often being excluded from consideration.

The Explicit-1 Phase

In the second phase the focus of learning shifts from external inputs to the learner's own representations. Internal dynamics take over and the current state of the learner's representations may predominate over the actual input. Representations at this level are not yet explicit in the usual sense, and thus the regularity in behaviour exhibited at this phase is what (Karmiloff-Smith & Inhelder, 1975) refer to as a *theory in action*, since it is still considered by the hypothesis to be somewhat embedded in action.

The formulation of a theory in action may result in new (and revealingly systematic) errors. If its predictions conflict with certain of the observable data, this may result in the pattern of decline (and subsequent recovery) of performance associated with a U-shaped behavioural curve. As Karmiloff-Smith (1992b, p. 20) points out, this intermediate deterioration is a 'deterioration at the behavioural level not the representational level' .

Although representations at this level are not available to conscious access, the use of the term 'explicit' here seems to derive some of its force from the idea that it is after this first redescription stage that manipulations, such as the violations of purely data-driven descriptions required for pretend-play, become possible, something which also suggests accessibility outside the original task mapping. Karmiloff-Smith hypothesises that such representations form the basis of a flexible cognitive system.

An important aspect of this phase is that it is hypothesised to begin only after an initial competence, or behavioural mastery, has been achieved at all or part of a task. This also has the implication that the onset of redescription is endogenous and is provoked by a period of stability rather than disorder in the organisation of what has been learnt.

The Explicit-1 Format

At level-E1, some of the sequential and procedural constraints acting at the first level are relaxed. At the same time some of the perceptual detail of the implicit representations is lost — E1 representations are often described by Karmiloff-Smith as being 'compressed' or 'reduced' versions of the original procedural and perceptual encoding.

This level of representation is also identified as being conceptual (where the previous level was characterised as procedural), in a similar sense to that used by Mandler (1988). Representations at this level also lose their 'bracketing' and become available for comparison with the

explicitised versions of other procedures, allowing the relationships between them to become 'marked and represented internally' (Karmiloff-Smith, 1992b).

Representations at this level take their place alongside the original procedural and perceptual representations, which are preserved in order that they can 'continue to be called for particular cognitive goals which require speed and automaticity' (Karmiloff-Smith (1992b), p. 21). The re-represented versions are then used wherever explicit knowledge is needed, for instance when components of the procedure are required for incorporation into another procedure.

Examples

In the microdomain of block-balancing (Karmiloff-Smith & Inhelder, 1975), it was found that children in the intermediate age-group (around 6 years old) were unable to balance blocks other than those whose weight was evenly distributed about their geometric centre. This is seen as evidence for the emergence of level-E1 representations. The hypothesis proposes that, through redescription of their implicit representations, these children have developed a theory-in-action based on the recognition of this regularity in the data, and that, since learning is now internally focused, this theory acts top-down and causes children to reject as unbalancable any cases which do not conform. This task setup also provides evidence for the claim that redescription preserves earlier, implicit, representations, in that it was possible, through an experimental manipulation, to force these original behaviours to be exhibited. When children who were acting according to the geometric-centre theory-in-action were asked to balance the blocks with their eyes closed they reverted to the proprioceptive strategy which characterised the implicit level and were again able to balance all blocks.

In the microdomain of the acquisition of possessive pronouns in French (Karmiloff-Smith, 1979b), after achieving an initial mastery in which words such as 'mes' (which conveys both plurality and possession) were deployed in a way corresponding to mature usage, children tended to pass through an intermediate phase in which each of these roles was marked explicitly and redundantly in the output using separate lexical items (e.g., 'toutes les miennes', 'all' + 'the' + 'my', for the usual 'mes' which implies all these functions in adult usage). Karmiloff-Smith (1979b) proposes that this change reflects a progression from representations of the roles of such words as a set of representationally adjunct 'unifunctional homonyms', learned under a strong innate one-form–one-function constraint, to a grasp of the plurifunctionality of individual lexical items. The initial attempt to reconcile plurifunctionality with the earlier constraint leads to the temporary overmarking of the separate functions.

Explicit overmarking of components is also seen as symptomatic of explicitised knowledge in the production of American Sign Language (ASL). Children learning ASL as a native language are seen to pass through a phase in which they mark out morphological components (which adult learners are unable to individuate) by (unconsciously) making previously fluid signs staccato.

Knowledge in E1 format is also thought to provide the control, flexibility and mobility of concepts needed for creativity. This knowledge is presumably not represented in either E2 or E3 format. As Boden (1990) notes, 'since not all aspects of skill are represented at a consciously accessible level, creative people usually cannot tell us how their novel ideas came about' (p. 73).

The Explicit-2 Phase and Format

Knowledge in E2 format is said to be explicit and available to conscious access but not verbalisable. In most presentations of RR to date, (e.g., Karmiloff-Smith (1992b)), no distinction is made between E2 and E3 pending further experimental investigation. The presence of E2 in the RR model is primarily intended not to exclude the possibility of a conscious, non-verbal format. Whether this format is a necessary precursor to verbal formats is an open question.

Examples

As an example of a case in which knowledge is consciously accessible but not verbalisable, Karmiloff-Smith (1992b, p. 22) cites the situations in which we are able to draw a diagram of something which we cannot express verbally. Karmiloff-Smith (1992b, p. 23) notes that '[n]o re-

search has thus far been directly focused on the E2 level ... most if not all metacognitive studies focus on verbal report'. An example of empirical work concentrating on this kind of representation is given by Goldin-Meadow and Alibali (1994). Their work has investigated the emergence of gestures symptomatic of representation in children performing Piagetian conservation tasks, before verbal report on those task representations was possible.

The Explicit-3 Phase

At the final level, knowledge is assumed to be accessible both to conscious reflection and to verbal expression. In domains in which intermediate performance is characterised by behavioural regression, this phase sees the reconciliation through redescription of the I and E1 representations into a single system capable both of correct performance and activities such as verbal reflection and analogy which require the later, more systematic representations.

The Explicit-3 Format

Karmiloff-Smith hypothesises that it is at this level that knowledge is recoded into a cross-system format, which is in addition 'close enough to natural language for easy translation into statable, communicable form.' (Karmiloff-Smith, 1992b, p. 23). She contrasts this idea with Fodor's claim that all knowledge is immediately transformed by innately-specified input modules into a common propositional language of thought.

It is hypothesised that for knowledge to be considered to be at this level, it must be able to support not only verbal expression but also certain other activities such as the making of explicit analogies.

Examples

Representations at level E3 are hypothesised to underlie verbal report. For instance in the experiments exploring children's comprehension of the article system in French (Karmiloff-Smith, 1979a), the oldest group were not only aware of the linguistic system they used to produce correct performance but were also able to discuss it. In particular they could now use their knowledge and awareness of the linguistic subsystem concerned in justifying their responses (younger children either could not justify these or referred to extralinguistic information), as well as using their knowledge to produce counterfactual examples to accompany these justifications.

E3 representations are also assumed to underlie activities requiring inter-domain accessibility of knowledge such as analogy, or those involving conscious exploration of knowledge such as explicit theory change (Karmiloff-Smith, 1992b, p. 16) and the formulation of real or conceptual experiments.

Although acknowledging that some knowledge enters the system verbally and is presumably thus stored directly in E3 format, Karmiloff-Smith argues that redescriptive processes are still relevant to such knowledge, since other knowledge in the same (or other) domain with which it interacts must be in the same cross-system format before this becomes possible. As Rutkowska (1993, p. 217) puts it, the E3 level of explicitation 'is thought to underwrite translation between different codes or systems of representation, in particular non-linguistic codes and language'. It is also noted that not all knowledge is redescribed into this format. For instance, Karmiloff-Smith (1992b) reports that long-range discourse constraints did not become available to conscious access or verbal expression.

The RR Process and the RR Model

Karmiloff-Smith (1992b) distinguishes between the RR model described above and the RR process which is what acts recurrently within the model. According to Karmiloff-Smith, if various features of the model, such as the ordering of phases, or the timing of behavioural mastery were shown to be in error, the validity of the process would not thereby be compromised. Evidence against the process would, on the other hand, constitute a challenge to the whole RR framework. Alternative RR-based models to that shown in figure 2.1 are possible (see Karmiloff-Smith (1992b, p. 24)), and indeed the current account of the RR model corresponds better to that in which E1 representations are redescribed into either E2 or E3.

In summary, representational redescription results in the existence in the mind of a set of multiple encodings of similar knowledge at different levels of explicitness. That these encodings form a conservative hierarchy is supported by the evidence presented by Karmiloff-Smith (1992b) that innate constraints as well as the theories-in-action resulting from explicitation are reflected in the structure of subsequent conscious explanations. Re-representations also form a hierarchy according to their accessibility beyond their original context.

2.3 The RRH in Context

The following sections put the RRH in context by comparing it to other theories of representational change and development, and by trying to establish its position on the key issue of representation.

2.3.1 The Position of the RRH

The RRH is offered as a speculative theory and a more or less implicit challenge to other theorists to provide mechanistic detail (Karmiloff-Smith, 1992b, 1994). However, as well as setting out certain constraints on the form of the RR model, Karmiloff-Smith also uses it to argue for general standpoints on development, discussed below.

Reconciling nativism and empiricism

The hypothesis is also characterised by a strong emphasis on the intention to reconcile nativist and empiricist theory. Karmiloff-Smith (1992b) thus defines her position in part by contrasting and relating it to those of both Piaget and Fodor. Briefly, she takes from Piaget a real role for development although allowing innate constraints more influence (see also section 2.5.1 below). She is very critical of Fodor's strong nativism, its attendant prespecified modularisation, and the notion that input undergoes direct translation into a common Language of Thought. However she does not reject the idea of modules wholesale, arguing instead that the products of learning (particularly in some domains such as language) are progressively encapsulated.

Also, in common with current trends in developmental theory, Karmiloff-Smith's approach incorporates aspects of both extreme positions in terms of the roles of innate and acquired knowledge. In Karmiloff-Smith (1992b), her position assumes the presence of innately-specified domain specific constraints in common with workers such as Spelke (1990) and Gelman (1990).

Causes of representational change

The RRH also involves a distinctive emphasis on endogenous change over exogenous change, although it does not exclude external influences as the cause of other representational change, nor denies them some sort of supporting role for RR. For instance, Karmiloff-Smith (1994, p. 738) claims that she has 'never argued that RR is *solely* generated endogenously'. She does consider, however, that RR is always provoked by positive feedback and stability rather than conflict and error.

In their response to Clark and Karmiloff-Smith (1993), Scutt and O'Hara (1993) contend that it is the pressure to make knowledge from one domain accessible in another which drives the process of redescription, rather than any stable representations spontaneously getting redescribed.

Other commentators have also argued that knowledge of external representational systems is inherent in redescription. Dennett (1993) argues that the ability to use linguistic labels contributes to redescriptive ability and conceptual mobility in general, as does pressure to express one's ideas via language. Olson (1994, p. 725) cites examples where knowledge of the written alphabet facilitates segmentation of words and phonemes, which he claims supports the idea that redescription is less a spontaneous appropriation of implicit features of the phonemic system itself than simply an adaptation of the categories offered by a cultural artifact such as the alphabet. Boden (1990) also notes the role of external notations in facilitating conceptual flexibility.

Karmiloff-Smith (1994, p. 738) responds that the RRH has never denied that literacy training during development affects brain configuration. She disputes, however, that literacy is necessary to awareness of word-boundaries, or for the phonological (as opposed to phonemic) awareness which sustains activities such as rhyming, as these abilities are found in illiterate adults and preliterate children respectively.

Conceptuality of early knowledge

There is some debate as to whether initial knowledge is conceptual or non-conceptual. A central proponent of the idea that knowledge in infancy is conceptual is Mandler (1988). Mandler proposes, that, in contrast with the Piagetian notion of sensorimotor schemas being transformed into conceptual systems, the two systems run separately and in parallel, with neither being derivative of the other. Mandler (1988) cites infants' capacity for imitation, recall of absent objects and motor recognition as evidence for the conceptual nature of early representations.

Although sympathetic to the evidence Mandler (1988) brings to her argument for the conceptual nature of initial knowledge, Karmiloff-Smith (1992b, pp. 77–78) does not actually make an explicit commitment to this position, although she is seen as implicitly adhering to a conceptual view of infant knowledge by Rutkowska (1994b) in her acceptance of central structures as mediators of behaviour. As Mandler (1988) acknowledges, the RRH, in proposing a series of shifts which link sensorimotor to conceptual knowledge, is not in precise agreement with her own view. However it does have some commonalities with the process by which Mandler claims infants encode perceptual information into an accessible system, and Karmiloff-Smith (1992b, p. 42) regards Mandler's work as a way of applying the RRH to infancy (see section 2.5.2 below).

The issue of representation in developmental study

The related issue of the representational status of acquired knowledge, (whether conceptual or non-conceptual, see Kirsh (1991) for instance) is a controversial one in both contemporary developmental studies (e.g., Smith and Jones (1993), Thelen and Smith (1994)) as well as in cognitive science (Brooks (1991), van Gelder (1992), Clark and Toribio (1994)). The debate concerns the extent to which internal representations are present, necessary and explanatory.

The traditional view of internal representation (see the account in Thelen and Smith (1994) for instance) sets it apart from action and knowledge embedded in specific instances of perceptual phenomena. Particularly in the case where these representations are also characterised as explicit or declarative (to be contrasted with the procedural representations at level I of the RR framework for instance), this division is regarded by Rutkowska (1993), for instance, as 'driving a misplaced wedge between knowledge and processes' (p. 127).

Smith and Jones (1993)'s argument for a holistic view of cognition cites the variability of cognition as grounds for rejecting conceptual accounts such as Mandler's. Their criticism of the traditionally central role of concepts is based on the claim that adaptiveness is central to intelligence, and thaand that it is not the stability of (generalised and abstracted) concepts which explains this adaptiveness but rather their variability and sensitivity to external input. They thus propose instead a pure-process approach similar to that of van Gelder (1992) in which concepts retain fluid perceptual cores.

Thelen and Smith (1994) are also critical of the objectivist Piagetian view of development as directed towards an end-point of transcendent, rational behaviour, and more specifically to the discussion of representation, reject the discontinuity not only of process but also of format:

> Language, logic, consciousness, imagination, and symbolic reasoning are not "above" the processes of motivated perception, categorization, and action ... [r]ather they are part and parcel of these processes, seamless in time and mechanism ... higher cognition is developmentally situated.
>
> (Thelen & Smith, 1994, p. 321)

A slightly different line of anti-representationalist criticism is pursued by Rutkowska. For instance Rutkowska (1994a) is doubtful that to the extent that robots use any kind of representation, this is likely to take the form of 'world-model'-style internal representation. She argues instead for a situated and action-based view of representation which acts to establish selective correspondences between subject and environment. According to Rutkowska the agent-specific bias inherent in notions of internal representation also makes it problematic as an interpretative tool, as it demarcates and also gives an artificial prominence to what is merely one component of a complex system in the genesis of organisation, thus limiting attempts to understand how that organisation is achieved.

An account of higher cognition without representation So what has the anti-representationalist literature to offer in terms of an account of phenomena such as metacognition and creativity most commonly assumed to require some form of conceptual or representational explanation? Thelen and Smith (1994) take up the challenge of accounting for higher cognition within a unified dynamical systems framework. As a way of integrating the idea of 'higher' cognition with their commitment to the idea of developmental continuity and situatedness they make use of Lakoff and Johnson (1980)'s idea of body metaphors. Briefly this is the notion that a great many familiar metaphorical linguistic constructs, which have come to seem quite abstract, originate in and, more importantly, remain grounded in, bodily experience. Thelen and Smith use the theory of neuronal group selection (TNGS) due to (Edelman, 1992) as a basis for discussing how this progression might come about: 'According to the TNGS, it is [the] continual forming and storing of varied categories that is the foundation for emergent higher-order abstractions.' (p. 325). More specifically, this process is characterised as a widening of influence as experiences originating in different developmental histories overlap where they have commonalities. This leads to something which seems as if it were a superordinate category which has emerged from these instances, but as Thelen and Smith (1994) reiterate '[t]his abstract knowledge ... is not a representation however, disconnected from its specific instances.' (p. 326).

This account also extends to the relationship between cognition and metacognition. Again, Thelen and Smith stress that 'higher' cognitive functioning remains grounded in action, for example:

> thinking about weaving will involve some of the same patterns of behavioural activity as weaving but one set will not be contained within the other nor will one be raised up to form the other.
>
> (Thelen & Smith, 1994, p. 337)

thus firmly identifying metacognition as a behaviour which self-organises from the real-time solutions of everyday life.

Mandler (1993) responds to Smith and Jones (1993) by arguing that the instability of perceptual categories has no necessary connection with the idea that perceptions remain at the core of concepts. She also considers that efficient deployment of perception does not explain the behaviour of children, who are able to form superordinate categories.

In his response to Brooks (1991), Kirsh (1991) also argues that representation, whether conceptual or non-conceptual (terms which correspond for Kirsh to the E2/3 and E1 levels of the RR model respectively), is necessary to an account of learning in that abstractions underlie the progressive transferability of knowledge.

The above account is suggestive, presenting a continuous, unitary mechanism within which action and metacognition may be accommodated. However it seems that the charge of explanatory poverty levelled at such accounts in the realm of higher cognition is not entirely dispelled by this rather impressionistic story, as its claim to be capturing metacognitive behaviours in any detail must still be seen as somewhat lacking.

Representationalism and the RRH It seems that Karmiloff-Smith seeks not to exclude entirely mechanisms from any particular cognitive scientific paradigm as potential explanatory mechanistic vehicles for the RRH. For instance, although concentrating on connectionist models she also alludes (in the closing pages of Karmiloff-Smith (1992b)) to the potential of the dynamical systems framework. Despite this apparent agnosticism, the explanatory metaphors in terms of which the RRH is presented give an indication of an affiliation with a position in the representationalism debate.

In particular, earlier accounts of the RRH make explicit use of the computer metaphor in characterising the interaction of the process of explicitation with representational formats. In Karmiloff-Smith (1990) for instance, she is explicit in her agreement with Rutkowska's view that action is best explained in terms of the concept of a program, specifically in the sense of that term which conveys something which can both be 'activated to generate processes' and, significantly, can be used as data by other procedures and manipulated. In the context of this analogy, part of the process of cognitive development becomes that which 'consists in building the second of these two functions, by redescribing the procedure at a higher level of abstraction such that the knowledge is then represented at two different levels.' (Karmiloff-Smith, 1990, p. 59).

Although the presentation of the RRH given in Karmiloff-Smith (1992b) is more explicitly in sympathy with aspects of a connectionist approach, terminology remains which is explicitly computer-metaphoric — the results of redescription are described as being 'abstractions in a higher-level language' (Karmiloff-Smith (1992b), p. 21), and also as data-structures. The use of such terminology suggests (at least) a commitment to the idea that external representational formats are an appropriate framework in terms of which to discuss internal representation and its change.

I suggest that there are two main problems with this. The first is methodological — the RRH is proposed as an avowedly diachronic hypothesis, conceived of with the intention of emphasising dynamics and the *process* of representational change over the representations themselves. While use of the computer metaphor clearly does not, of itself, exclude useful discussion of process, as the above examples show, Karmiloff-Smith tends to concentrate on its static aspects such as data structures and notations in order to explain representational formats. Even this is problematic in that, to the extent that programming notations form a sort of continuum, the gaps between any two instances tend to be larger than seems to suit the explanatory grain of the RRH at the microdevelopmental level, and differences are often unhelpfully qualitative. In particular, the analogy to programs does not provide us with any way of elaborating on the progression from one status to the other in the way that I will argue process models such as connectionist or dynamical systems models do. This is because, without bringing the human programmer into the picture, there are few examples in computing of automatic processes (disassemblers being a notable exception) which act to transform notations into others which are qualitatively more abstract, and it seems likely that this scarcity of examples may explain this focus on structure. Although automatic symbolic abstraction procedures exist, such as that used in the lambda calculus, this does not seem to be the kind of radical qualitative transformation envisaged by the RRH, but corresponds better to the declarative–declarative transformation possible at E3.

While the above indicates that Karmiloff-Smith regards representational constructs as useful explanatory metaphors, it does not commit her to the idea that representations are necessary. But in Karmiloff-Smith (1992b)'s discussion of what might constitute criteria for disproof of the RRH, clearer indications emerge of her affiliation to a particular view on the status of internal representations are to be found. In her view the process of representational redescription would lose plausibility if, for instance:

> all representations in the mind were of equivalent status, or if totally distinct constraints were operative on procedural versus declarative knowledge, rather than each level involving redescription of the previous one.

(Karmiloff-Smith, 1992b, p. 25)

This strongly suggests a position which is in opposition to the continuous, same-status view of Thelen and Smith (1994) for instance, and in a brief review of that work (Karmiloff-Smith & Johnson, 1994), her position is that while Thelen and Smith are to be congratulated for both the dynamic approach taken by their work and their rejection of the computer metaphor, their denial of a 'special status' (p. 53) to representations in the brain (as opposed to other physical sources of control variables) is misguided. Karmiloff-Smith and Johnson's main justification for this view is that even if conceptual categories emerge from the interaction of multiple levels of information, 'they are still qualitatively different from the perceptual categories.' (p. 54).

And in turn, Thelen and Smith (1994) criticise the RRH specifically for its apparent commitment to the idea that concepts have a 'transcendent' relationship to lower-level thought. Redescription on this view 'raises up' more basic processes to give a layer above real brain activity, rather than being a product of global activity. This idea, in their view, raises many problems, including a suggestion that redescription implies a homunculus, and the problem of how symbolic thought is to be considered as being represented if not also as distributed patterns of activity over time within the same dynamic system as the original activity.

Although some presentations of the RRH (e.g., Karmiloff-Smith (1990, p. 77)) do assume a notion of central processing, in which representations come to be represented explicitly, the fact that the redescriptive process is postulated to be domain-general need not imply that it acts centrally. Also, the implementational suggestions (e.g., connectionism) made by Karmiloff-Smith (1992b) suggest that redescription could be sustained by a locally acting process of self-organisation similar to that proposed by Thelen and Smith (1994).

In the sense that redescriptions, at whatever level, are part of a single hierarchical system, there is no implication that higher-level representational formats require a different implementational substrate.

Another criticism, which is closer to the core of the idea of the RR model, is that although the idea of hierarchy and the retention of grounding concepts and processes is inherent in the model, there is no suggestion that higher-concepts remain grounded in these lower ones. As Thelen and Smith observe, the relationship between higher and lower cognition only flows one way — lower concepts contribute nothing to higher concepts once formed and are mainly preserved for situations requiring efficient performance.

Despite the obvious conflicts between the RR model and non-representational accounts of development, it seems that particularly in terms of implementations and potential mechanisms the two approaches could be reconciled in terms of dynamic self-organising environmentally grounded knowledge programs, in the way Rutkowska (1994b) advocates. The reformulations of key concepts, such as 'theory' in terms of process, by Clark (1993a), also go some way towards an RR model which is distanced from the idea (which Clark (1993a, p. 81) finds too rationalistic) of RR as involving reflection on previously acquired knowledge. These reformulations also constitute a way of operationalising notions such as explicitness, and form some of the central working assumptions in this project.

2.4 The Scope of the RRH

This section surveys the main issues raised in debates concerning the applicability of the RRH, and in doing so tries to map out where its predictive boundaries might lie.

2.4.1 Ontogenetic boundaries

The empirical work out of which the RRH arises has been conducted primarily with subjects in middle-childhood (specifically between the ages of 4 and 11, see Karmiloff-Smith (1990), Karmiloff-Smith and Inhelder (1975), Karmiloff-Smith (1979b) for instance). However Karmiloff-Smith (1992b) also attempts to link infancy into the RR framework, and further claims that the

hypothesis can be used to account for representational change in adult learning — albeit only in certain domains, specifically those (unlike language in particular) in which knowledge has not become encapsulated through the process of progressive modularisation, which is assumed to accompany redescription. I consider each of these in turn.

Infancy

As Karmiloff-Smith acknowledges, the RRH stems from work on subjects in middle childhood and initially made no attempt to take infancy results into account. Karmiloff-Smith (1992b) however, cites the volume of recent work on infancy as a primary motivation for including it in discussion of the RRH. According to Karmiloff-Smith, the main consequences of this new attempt to integrate infancy are to be seen in the epistemological framework, which this work tries to establish, of a reconciliation between nativism and constructivism, and more specifically in the highlighting of domain-specific constraints on development.

Despite the new prominence given to domain-specific (and usually innate) constraints in the presentation of the RRH in Karmiloff-Smith (1992b), it is also claimed that '[a]s a model of representational change, it would stand unaltered even if it turned out that there were no innate predispositions or domain-specific constraints on development' (p. 165). Karmiloff-Smith's primary interest in infancy in the context of the RRH is the representational status of infant knowledge. It is claimed that, in the framework of the RRH, it would probably be inconsistent to regard this knowledge as a 'theory' as, for instance, Spelke does, since the hypothesis requires that knowledge be represented in at least E1 format before it has this status. Infant behaviours on the other hand often seem to require no more than representation in I-level format. Specifically, Karmiloff-Smith prefers to characterise infant knowledge as procedurally represented (see Rutkowska (1993)), in the sense that, while not seeking to deny that infant knowledge is both rich and coherently organised, she also contends that it is 'first *used* by the infant to respond appropriately to external stimuli' (Karmiloff-Smith (1992b), p. 78). This gives it a procedural representational status and suggests its integration into the RR model at the I level.

In terms of the RR model, Rutkowska concurs with this, in that she does not consider the conscious explicit formats (E2 and E3) to have particular relevance to an account of infancy, believing instead that '[o]verall, the three-phase model [1] current I and E1-levels appear to provide an appropriate space within which to locate the intrasystem representational changes needed to account for infancy.'(p. 217), and giving examples of how the three-phase model seems to correspond to the recurring three levels of infant performance. The idea of a three-phase progression in infant behaviour is also central to the work of Mounoud (e.g., Mounoud (1982)). Rutkowska (1994b) is more critical of Karmiloff-Smith, referring more specifically to the problems inherent in Karmiloff-Smith's assumption that infant knowledge (both innate constraints as well as that arising through learning) is characterised only by the implicit format. Rutkowska (1994b) is doubtful about this mapping, commenting that '[s]ince level-I representations are limited to mediating the context-bound input-output relations that underlie behavioural mastery at least the E1 level might be expected.' (p. 727) since properties such as systematicity and predictability would seem to be necessary to a characterisation of knowledge at this level, and in these, the (minimal) notion of explicitness which E1 would seem to be intended to embody.

The difficulties encountered by Karmiloff-Smith in mapping the formats of the RR model (including Mandler (1988)'s notion of an image-schematic format which seems to lie between the implicit and E1 formats) onto infant knowledge, are cited by Rutkowska (1994b) as grounds for rejecting not only, as does Karmiloff-Smith, that infants have theories, but also that their knowledge is conceptual in any sense usually employed in philosophy. Rutkowska suggests instead that by viewing the infant as a situated agent and regarding any central processing as not orientated around the fixation of propositional beliefs but the flexible coordination of perceptual and

[1] An earlier formulation of RR model in which representations progressed through 'procedural', 'meta-procedural' and 'conceptual' phases (see Karmiloff-Smith (1984) for instance), and which Rutkowska (1993) considers to correspond more closely with phases in infant performance.

behavioural components in a supporting environment.

Other issues also remain to be addressed. For instance what precludes infants from redescribing this knowledge beyond level-I in the way that older children do? Rutkowska (1993) points out that the path to initial mastery is not addressed. Karmiloff-Smith also seems to say nothing about the timing of any initial redescription, or about the possible age of onset of the domain-general predisposition to redescribe or what provokes this.

The integration of infancy into the RR framework, although providing epistemological grounding and evidence for the conservativeness of the redescriptive process, is still a partial one in the account provided by Karmiloff-Smith (1992b), in terms of the questions of timing and formats that it leaves open. It seems that infancy studies still have more to offer the RRH than it has to offer in return.

Adulthood

The RRH makes two main claims regarding adult learning. The first is that redescriptive effects, in particular conscious reflection, are associated with the development of scientific theories and exploration of analogy (Karmiloff-Smith & Inhelder, 1975). Although these activities tend to be restricted to adults (and older children) it is hypothesised that they are 'possible only on the basis of prior representational redescription' (Karmiloff-Smith, 1992b, p. 16).

The second claim is that redescription can apply to earlier, non-conscious parts of the adult learning process as it does in children. As evidence for this Karmiloff-Smith (1990) notes that the phonological representations of newly literate adults are subject to sequential constraints which suggest that a phase of redescription has occurred. She goes on to infer that 'the [RR] process involves a phase in a reiterated cycle of representational change, and not a developmental stage only to be found in children.' (Karmiloff-Smith, 1990, p. 78).

The main constraint given on the redescription of adult learning is the extent to which the domain has become modularised, as such knowledge is no longer available for redescription. It is considered likely that language is such a domain. For instance, Karmiloff-Smith (1992b, p. 49) reports that deaf parents learning American Sign Language as adults do not go on, as children do, to analyse the signs' morphological structure. Despite the E1-level redescription in the case of the newly literate adults above, there is no indication given that their representations progress beyond this stage and become consciously accessible or verbalisable.

Further investigation is clearly necessary to establish the constraints on this progressive encapsulation across domains, as well as that of the details of conscious conceptual exploration and its place in the RRH.

Micro- and macro-developmental change

The RRH is presented both as a theory of development in which the characteristic phases and representational changes occur over the course of as much as several years, as well as one of task-learning in which a similar pattern may emerge over the course of a single experimental session (Karmiloff-Smith, 1979b, 1992b). Karmiloff-Smith (1979b) relates the two levels (as observed in experiments on plurifunctionality of linguistic forms and map drawing) thus:

> In both cases, an initial phase of superficially complex forms is followed by a phase during which children indicate by concrete external markers each piece of information they wish to convey.
>
> (Karmiloff-Smith, 1979b, p. 114)

So what is the 'default' level of granularity at which the RRH has explanatory power, and how are the two levels of granularity related? Karmiloff-Smith (1992b) suggests that the macro-developmental level is the default level of explanation of the RRH:

> in previous chapters ... it was established that representational change does indeed occur macrodevelopmentally. Here I address microdevelopmental change, i.e., change that occurs within the confines of an experimental session.

(Karmiloff-Smith, 1992b, p. 148)

while the following suggests that microdevelopmental change has something of a secondary role:

> If, as I argue, representational change is pervasive in human development, then there is no a priori reason to limit it to the macrodevelopmental time scale. It should be possible to establish its occurrence also in the microdevelopmental timescale.

But although Karmiloff-Smith stresses the similarity of the processes at work on these different timescales, there is little indication of how they are related — specifically for us as computational modellers, how microlevel changes accumulate, contribute to and bring about macrolevel changes. The implication that RR has a self-similar microstructure, or simply that it *can* provide explanations of phenomena at different levels, according to the domain, remains an open issue.

In the map-annotation task described in Karmiloff-Smith (1979b, 1992b), the introduction of redundant information into subjects' notation is explained as the explicit marking of information implicit in the original, efficient, system.

So, despite the overall similarity, can we pick out significant differences between the form of the RR model at the micro- and macro-levels? The most obvious difference seems to lie in the proposed level of accessibility of the final phase of RR. In the macro-level model this level is explicitly designated E3 and associated both with verbal expression and conscious awareness. Although at the microlevel the final RR phase is seen to follow a similar pattern of reconciliation between initial implicit and subsequent explicitised knowledge formats, this is not associated with a specific level or type of access.

Another difference is that the solution of the task must be well within the subjects' competence if redescription is to be observed over the course of a task (Karmiloff-Smith takes this as providing evidence for the necessity of prior behavioural mastery for redescription).

2.4.2 Phylogenetic boundaries

Karmiloff-Smith's attitude towards possible phylogenetic breakpoints in RR has also undergone some changes during the evolution of the RRH. There are repeated suggestions that redescription is a distinctive feature of human cognition and, if not, is at least likely to be far less spontaneous and widespread in other animals (Karmiloff-Smith, 1979b, 1990, 1992b)

However Karmiloff-Smith (1992b, p. 16) also adds that she considers explicit theory change and the other conscious activities which characterise level E3, to be 'more obviously restricted to the human species'.

But although there is a difference in general competence between even chimpanzees and children after quite an early age, there is some evidence against a sharp phylogenetic breakpoint in capabilities associated with redescription and metacognition in general.

Some controversial evidence from recent studies with chimp and gorilla sign language (e.g., Patterson, Patterson, and Brentari (1987)) suggest that apes are capable of puns and 'rhymes' — activities which would seem to require both access to component morphemes, and which suggest something of the kind of spontaneous tendency to treat one's acquired linguistic knowledge as a system that the RRH demands. Boysen, Berntson, Shreyer, and Hannan (1995) found that chimpanzees experienced in counting small arrays and comprehension of number symbols spontaneously displayed gestures such as pointing and rearranging items, indicating the structure of their representations of number in a way similar to that observed in young children.

It seems likely that the enculturation of apes, specifically the teaching of symbolic systems of communication (e.g., Savage-Rumbaugh, Murphy, Sevcik, et al. (1993), Boysen and Berntson (1995), Boysen et al. (1995)), affects their tendency or ability to form structural, transportable conceptualisations whose components they can then manipulate. For instance, Boysen and

Berntson (1995) found that chimpanzees who had been trained to use Arabic number symbols were able to overcome perceptual-motivational cues in a task requiring them to choose the smaller of two piles of sweets only when numerals were used instead of the sweets themselves.

Karmiloff-Smith (1983) speculates that such systems may provide them with the necessary abstract code, or the possibility of forming multiple representations. This has implications for the issue of the role of external notations for redescription in humans discussed above, and also suggests that redescription, or its ingredients, may well be as domain-specific in non-human animals as they are in humans.

Donald (1994) suggests an explicit role for the succession of increasingly explicit formats of RR as phylogenetic intermediaries — although arguably in characterising human cognition in terms of E3 representations he is omitting the earlier, more procedural levels of representations in human learning.

Although the place of animal cognition in the overall picture of representational redescription remains to be determined, it seems probable that such a tendency to work on one's own knowledge is significantly more pronounced in humans, even if it is present in non-human animals. However it would be interesting to investigate whether a similar pattern of redescription-like phases was at work in non-human primates' learning of language or number as discussed for the case of infants in section 2.4.1 above.

2.5 The RRH and other theories of the emergence of explicit knowledge

2.5.1 Piaget

Although Piaget's theory of qualitative developmental change and the emergence of abstract thought has broad similarities with the RRH, there are important differences which Karmiloff-Smith stresses. Piaget's is a *stage theory* (see Piaget (1953)), proposing across-the-board change, whereas the RRH posits domain-specific phases of redescription.

Karmiloff-Smith is in support of the epigenetic perspective on knowledge and development, but believes that innate constraints play an important role in guiding development and in this her viewpoint differs strongly from Piaget's.

The endogenous nature of change in the RRH also differentiates it from Piagetian theory. For Piaget, qualitative change, via the process of *equilibration*, occurs in response to a state of disequilibrium caused by external information which is beyond the scope of the system as it stands. While not denying a role to externally driven change, the RR model focuses on change which comes about after a period of comparative stability has been reached.

2.5.2 Mandler

Mandler (1988, 1992) gives an account of early knowledge in which a process of *perceptual analysis* acts to render perceptions into first an image-schematic format and then subsequently into linguistic form. Mandler (1992, p. 589) refers to this process as a simple version of the redescription of procedural information found in the RRH, and an obvious parallel with the RRH can be seen in this progression through formats. Karmiloff-Smith (1992b) relates Mandler's work to the RRH thus:

> The redescription of perceptual primitives into image-schematic representations and of the latter into language, indicates how the RR model ... can be applied to very early infancy. I have stressed the fact that representational redescription can occur outside input–output relations, Mandler extends the RR model to on-line processing, suggesting that redescription also takes place as the child is actively engaged in analyzing perceptual input and redescribing it into the more accessible format of image schemas
>
> (Karmiloff-Smith, 1992b, p. 42)

Mandler also has it that some detailed information is lost through perceptual analysis, as in the RR process, and that it is based on an innately specified analytical mechanism, which may however act on innate or acquired knowledge.

2.5.3 Halford

Halford (1993) describes a theory of cognitive development based on representations (mental models) which increase in the dimensionality of the relations they can capture over time.

Halford also differentiates between implicit and explicit representations on the basis of cognitive accessibility. He suggests that implicit representations in the form of contingencies are recoded into explicit condition–action pairings, whose components are cognitively accessible. This recoding involves a process of abduction, i.e., of forming a hypothesis about a contingency through reflection upon it.

Halford (1993, p. 50) is also in agreement with Karmiloff-Smith that the ability to transfer knowledge to isomorphic tasks and to reorganise the relational structure of domains in order to relate them to other domains are requirements for a definition of understanding.

2.5.4 Relating the RRH to a general associative–relational divide

Several workers in cognitive science have related the RRH to a more general account of implicit and explicit thought, which aligns them with associative and relational learning and representation. For instance, Philips, Halford, and Wilson (submitted), working in a framework similar to that of Halford (1993), link implicit to associative and explicit to relational knowledge respectively, on the basis that the latter is *omnidirectional*, i.e., any component is accessible via any other and the roles of these components are individuable.

Thornton (1995) makes a similar alignment. According to his account, implicit knowledge is that which is embodied in relations (as opposed to being manifest in the statistics of the input data), and the work of an explicitation process is to bring such knowledge within the grasp of an associative learning mechanism.

Both of these formulations differ from the RRH as presented by Karmiloff-Smith in being purely domain-general, in common with many machine learning algorithms, (although in the implementational work of Thornton (1995) (see chapter 3) task-specific biasing is used to scaffold learning).

2.6 Domain-specific differences and the RR model

Despite the emphasis on domain-specificity in presentations of the RRH such as Karmiloff-Smith (1992b), the RR process is hypothesised to be domain general. However some of the difficulty of characterising RR in general derives from the fact that it manifests itself in differing ways in different domains.

2.6.1 Behavioural marking of E1 representations

In lexical morphology (Karmiloff-Smith, 1979a, 1979b), a U-shaped behavioural curve in conjunction with overmarking is used to diagnose redescription to E1 format. In the block balancing task of Karmiloff-Smith and Inhelder (1975), the emergence of explicit representations is also marked by a decline in performance as well as systematic errors reflecting a theory-in-action.

In the domains of counting and music (Karmiloff-Smith, 1992b) however, no such macro-developmental U-shaped curves or external behavioural marking is reported. Karmiloff-Smith acknowledges that behavioural marking is not necessary to a diagnosis of redescription.

2.6.2 Conservation of earlier representations and procedures

Karmiloff-Smith stresses the fact that redescription is not a drive for economy (Karmiloff-Smith, 1992b, p. 23), rejecting analogies with data compression or garbage collection[2] — representations are, rather, conservative and hierarchical.

Part of the evidence for this is provided by the ability to elicit an earlier (and more successful) strategy from children in the block balancing task. The RRH has it that the level-I procedures (here balancing blocks using proprioceptive feedback) are preserved for use in efficient production.

But is this always the case and does it apply to representations at the higher, explicit levels? It would seem rather odd to categorise the presence or absence of an effect which is proposed as central to RR as a domain-specific difference.

For instance, in the domain of lexical morphology, it does not seem to be the case that the earlier unifunctional homonyms are preserved as such, although the phonological procedures to produce the words may be. The idea of a change in status here seems to imply that these are reappropriated more radically. It would be interesting to see whether an experimental manipulation exists which would provoke a return to the earlier stage in older children or adults.

From the evidence surveyed in Karmiloff-Smith (1992b) for instance, it is also difficult to see that aspects of E1 or E2 representations are preserved in the same way in the redescribed E3 format. In the block-balancing task, the I-level theory in action is reflected in subsequent representations. If this effect were observed across a number of domains it might violate the idea that RR is conservative and hierarchical at all levels.

2.6.3 Extent of redescription

As Karmiloff-Smith acknowledges, redescription need not reach level E2/3. Karmiloff-Smith (1979b, p. 97) also reports a case in which the behavioural symptoms of the three phases are observed but without verbal or conscious access having been achieved. Karmiloff-Smith (1994) acknowledges Scholnick (1994)'s observation that the RR model lacks a principled way of discriminating between domains which do or do not become modularised. Karmiloff-Smith suggests that these differences may be due to competition for computational resources.

2.7 Other responses to the RRH

This section surveys general responses to the RRH itself. Responses to implementational proposals made by Karmiloff-Smith and her collaborators (see Clark and Karmiloff-Smith (1993), Karmiloff-Smith (1992b, 1992c)) are discussed in chapter 3 below.

2.7.1 Form of the RR model

Issues raised in this area can be divided into two main categories. Commentators who lack a basic sympathy with the idea of representational format which the RRH puts forward have tended to direct their criticisms towards the nature of formats in the RRH, while others focus more on issues affecting the structure of the model at a more macroscopic level, such as the number and sequencing of formats.

Number of representational formats

Carassa and Tirassa (1994) put forward the general concern that proposing many representational formats entails also proposing a large amount of detecting and decoding machinery. Goldin-Meadow and Alibali (1994) provide experimental support for Karmiloff-Smith's four-format story. Evidence for representations at Karmiloff-Smith's level E2 comes from work in which conscious awareness is revealed through gesture before verbal access has been gained.

[2] the automatic periodic removal of data-structures no longer needed by a computer program in order to save space

Many levels vs. simple implicit–explicit distinction

de Gelder (1994) uses evidence from the domain of language to argue that implicit and explicit systems can dissociate. In Donald's evolutionary account, (Donald, 1994), the two paths which he claims have evolved for access to implicit memory seem to take knowledge directly from I to (either or both of the) E2 and E3 formats, with E1 having a role perhaps only as a phylogenetic intermediary in the development of fully explicit representations in humans.

Sequencing of representational formats

de Gelder and Carassa and Tirassa are worried about the kind of 'temporal logic' assumed to link implicit to explicit representations in the RRH. Carassa and Tirassa (1994) make the point that the fact that procedures are learnt first need not mean that initial knowledge is procedurally represented, and that some knowledge starts off in declarative form, a point which Karmiloff-Smith (1992b) acknowledges.

Goldin-Meadow and Alibali (1994) claim that studies of gesture suggest that accessibility (and indeed redescription) may require not mastery as the RRH proposes, but merely stability. According to the account of the conditions under which the RR model might be refuted as set out by Karmiloff-Smith (1992b) (pp. 23–25), this has implications for the validity of the model.

Peterson (1993) examines and rejects the RRH as a potential theory of general re-representation, explicitly avoiding discussion of its status as a theory of cognitive development (p. 3). In particular he is concerned with the kind of declarative–declarative transformations of problem formulations that characterise conscious adult problem solving. He argues that in the examples given, re-representations of the problem domain lead not to 'more succinct *statements* about a domain' (p. 3) as the RRH might suggest but to improvements in procedural performance. I would argue that there is nothing in the RRH to suggest that redescription cannot result in improvements in performance; it is simply that the need to make such improvements does not provoke redescription. Also, Karmiloff-Smith claims that explicit problem transformation, for instance using analogy, is facilitated by the products of previous redescription, and involves manipulations on declarative representations, just as Peterson suggests.

Sequencing of accessibility

Scholnick (1994) considers that the processes which must underlie the initial implicit–explicit transition differ radically from those which transform the resulting explicit representations into verbalisable form.

2.7.2 Nature of representational formats

Campbell (1994), Rutkowska (1994b) and Vinter and Perruchet (1994) are all unhappy about the epistemological status of representational format in the RRH. For Vinter and Perruchet (1994), even initial mastery may well have to be underlain by explicit knowledge, since there is evidence to suggest that implicit knowledge may not contain embedded knowledge of rules for later reappropriation and explicitation. Rutkowska (1994b) is unhappy about the suggestion that implicit knowledge could be conceptual, a claim made more explicitly by Mandler (1988), while Campbell (1994) argues that formats cannot be viewed as encodings in a formal sense since this leads to a lack of basic grounding for the representations.

Another set of commentators focus on the idea that representational formats are insufficient in themselves to explain increased accessibility. Olson (1994) stresses the importance of explicit external categories in the explicitation process. Boden (1990) highlights the role of language, including technical forms such as music notation, in supporting the passage from domain-specific structure to conscious access.

For Losonsky (1994), a 'procedural representation' relies on the ordered and integrated deployment of both internal and external representations via some sort of feedback loop.

Carassa and Tirassa (1994) also claim that it is the content of the representations which determines their accessibility rather than anything in their format, and further that RR seems to

deal with how knowledge contents are used rather than how they are represented.

Formats and verbalisability

The RRH claims that representations must be redescribed into E3 format before they become accessible to both consciousness and verbal expression. This format is also seen as facilitating transferability of knowledge outside the original domain.

This link is criticised by Carassa and Tirassa (1994), who point out that even Language of Thought theories do not require knowledge to be represented in a form similar to natural language in any respect but constituency, and that this in itself is sufficient for interdomain transfer, and in a way which does not needlessly exclude members of most other species.

For Donald (1994), the RRH is wrong in proposing separate formats for different kinds of explicit access — for him, the important thing about a format is that it supports explicit access whether verbal or non-verbal.

Origins of knowledge and redescription

Bloom and Wynn (1994) are concerned that whereas, as they see it, the real challenge to any constructivist developmental theory is to explain how knowledge arises from development, the RRH concentrates on how knowledge which is already in the mind is redescribed. Campbell (1994) also worries that RR cannot introduce new knowledge into the system.

Several commentators raise similar issues in relation to the 'takeoff' of the RR process itself. For instance, Olson (1994) considers that RR must be externally driven since the individuation of components it brings about must rely on the provision of external explicit categories for instance via cultural artifacts. Scholnick (1994) makes a similar point in attributing a socio-cultural origin to redescriptive 'skills' and knowledge to input.

Grounding of representations

Campbell (1994) raises doubts about the epistemological status of representations; if RR formats are 'encodings', then these need to be grounded and I-level representations are not the 'foundational encodings' necessary.

2.7.3 Motivations for redescription

The E1–E2/3 transition

Some commentators are specifically concerned with what motivates the final transition from phase 2 to phase 3 (formats E1 to E2/3) in micro-domains which exhibit U-shaped curves. For Dennett (1993), it is the pressure to verbalise which drives all RR, while for Scutt and O'Hara (1993), it is the continuing external pressure to perform which drives the learner to regain performance, rather than internally driven processes.

RRH as a general theory of re-representation

Peterson (1993) takes a problem-orientated view of re-description. In his example problems the redescriptions proposed are closer to the shifts in viewpoint discussed in the knowledge-representation literature (for instance the re-representation of the 'missionaries and cannibals' problem as a graph of legal states and transitions) and seem to be in direct contrast with the progressive procedural–declarative redescription of the RRH.

This fact is at the heart of Peterson's criticisms of the RRH; for him the account of procedural–declarative representations is too narrow to capture declarative–declarative problem re-representations and is thus incomplete as a theory of representational redescription in general.

He bases his comparisons on a characterisation of Karmiloff-Smith's position in terms of the following five issues: the existence of ontogenetic and phylogenetic breakpoints, spontaneous endogenously driven change, abstraction, procedural-to-declarative transformations and implicit-to-explicit transformations. He then tries to show that these characteristics do not apply to all cases where representations are redescribed, by comparing the re-representations from the knowledge-representation literature with the list of characteristics. The redescription of a

game called number scrabble as a game of noughts and crosses over a magic square, and the Roman and Arabic numeral systems are presented as examples and Peterson makes the following analysis of the applicability of his list of characteristics. Although he is uncertain as to whether such redescriptions can be termed abstractions, his criticisms focus on the nature of the transformations involved. In number scrabble, he argues, the transformation is not from procedural to declarative, but rather from procedural to procedural, the virtues of the re-representation being to reduce the load on working memory rather than to increase the expressive power over a domain, while in the case of numerals, the point of re-representation is to facilitate arithmetic procedures. In terms of explicitation, he argues that the transformations involved in number scrabble are from explicit to implicit since knowledge becomes incorporated in the diagram itself, while in the numeral case no previously implicit knowledge is rendered explicit by the act of re-representation.

Although Peterson seems to focus his criticisms on the initial implicit–explicit level of redescription, it is arguable that the examples he gives would be characterised by the RRH as involving purely procedural–declarative transitions as he implies.

The notion of re-representation he employs seems to be more in keeping with that which might occur at the E-2/3 level, at which knowledge is seen as explicitly represented and transferable between tasks. There is no suggestion that the RRH seeks to exclude the possibility of declarative–declarative redescriptions especially at this high level.

It is also arguable that the example in which the Arabic and Roman numeral systems are compared is actually an example of redescription at all, although both are systems for representing the same data. It may not even qualify as an example of re-representation since this would seem to imply that one system was devised as a transformation of the other rather than both having emerged independently from the need to represent large quantities.

Peterson also stresses the 'point' of redescription in these examples ('to facilitate arithmetic', 'to reduce cognitive workload' for instance), seeming to suggest that such goals place his example problems outside the space described by the RRH. However there seems no suggestion that such facilitation is outside the predictive scope of the RRH, even if some of the declarative representations formed may cause temporary declines in performance. Indeed in Rutkowska (1993)'s view '[t]he goal of this procedural reorganisation is greater control over the environment and action on it' with representations seen as mediating procedure and representational redescription as a means of developing the ability to use knowledge to anticipate and form pre-conditions for actions.

Peterson's work may be best seen as an examination of the relationships which exist between differing representations of the same problem, and interesting questions remain as to how 'redescription' in the sense he intends it may be linked to the term as used by Karmiloff-Smith — in particular, such redescriptions may well be seen as characteristic of the RRH at E3. However it is not obvious that he is comparing these two formulations of redescription at a similar level.

2.8 Modelling the RRH

Although the enterprise of constructing a computational model of the RRH is the central focus of this thesis, it is appropriate to make some general comments here about the motivations for such an undertaking and the possible mechanisms for the RRH.

The speculative nature of the RRH

According to Karmiloff-Smith (1994) the RRH is formulated as a speculative theory in the anticipation that precise mechanisms will be provided by others with expertise in modelling.

Karmiloff-Smith (1992b) cites Klahr's distinction between soft-core and hard-core approaches to modelling development. This distinction is basically that between models which focus on the general processes at work or on specific mechanisms. Karmiloff-Smith is eager to emphasise the

complementary nature of these approaches and defends soft-core approaches such as the RRH on the basis that they avoid premature commitment to artificial or terminological separations between processes which are in fact fluid or interactive. In her view soft-core approaches thus support a better general conception of processes.

Motivations for the computational modelling of development

General motivations for constructing computational models for developmental phenomena include the fact that, as Klahr (1995) argues, irrespective of paradigm, computational models (in particular, so-called *process models*) offer theorists a chance to examine their hypotheses under dynamic conditions. This process may then expose weaknesses which were not apparent from the original static formulations of a particular theory.

Rutkowska (1993, pp. 3–6) however is skeptical of the intrinsic value of ad hoc translations of developmental principles into programs in traditional AI languages such as LISP and Prolog, and cautions modellers to focus instead on models of proven worth which 'illustrate *robust* ideas from [cognitive science] about the way computation might be organized' (p. 4).

Exploring constraints on redescription

Another motivation for modelling cited by Karmiloff-Smith (1992b) is to try to discover the constraints on the process of redescription itself. This might in turn provide answers to questions such as the status of redescription which does not lead to verbalisability for instance (Karmiloff-Smith, 1992b, p. 188), or redescription in adults versus that occurring during development. It might also provide explanations for why certain features of a domain are redescribed and in what order this must occur.

Other issues, such as the role of domain-specific and domain-general constraints, the default level of RR, the relationship between explicit and accessible representations, and the roles of external input or continued on-line processing, are amongst those which a computational model might help to clarify. On the other hand it seems less likely that computational modelling as it stands could tell us so much about such comparative questions.

Representational redescription and connectionism

Connectionism in particular has been linked with the RRH (Clark & Karmiloff-Smith, 1993; Karmiloff-Smith, 1992b, 1992c) on the basis that, of current computational modelling paradigms, it seems 'closest to the spirit of epigenesis and constructivism' (Karmiloff-Smith, 1992b, p. 176), and secondly that 'a number of features of the RR model ... map interestingly onto features of recent connectionist simulations' (Karmiloff-Smith, 1992b, p. 176). In particular, the path to behavioural mastery seems to correspond well to the gradual adjunction of representations leading to stability which is characteristic of learning using gradient-descent methods such as backpropagation.

Computational models of development, the RRH, and connectionism in particular will be the focus of the next chapter.

2.9 Summary

In this chapter I have introduced the representational redescription hypothesis and described the RR model in detail, presenting examples of each phase and its associated representational format. The hypothesis was put into context through comparison with other developmental theories, in particular those of Piaget, Mandler, and Halford. The applicability of the hypothesis to infants, adults and non-human animals was discussed.

In terms of adherence to the idea of strong internal representations, it was argued that there was a conflict between Karmiloff-Smith's talk of dynamical systems and the use of computer-metaphoric terminology in the presentation of the RRH. Although representation is considered

necessary to the RR model, it was argued that in the implementational suggestions given by Karmiloff-Smith (1992b), the suggestions of Rutkowska (1993, 1994b), and the reformulation of the RRH along connectionist lines by Clark (1993a), certain aspects of the dynamical systems perspective might be reconciled with the RRH, in particular the notion of different representational format as gradual increments in multiple usability.

The predictive scope, although touching on infancy and adulthood, was still found to centre on middle-childhood, while suggestions that redescriptive processes occur in non-human animals are still very much open to debate.

Criticisms of the hypothesis centre on the form of the RR model, in particular its discontinuous and conceptual representational formats, and the strain evident in the attempt to apply it to infancy. The RR process itself is less critically received (perhaps partially because it is described in much less detail).

Motivations for constructing a computational model of the RRH include providing, and testing dynamically, candidate mechanisms for the RR process or model, and thereby also investigating constraints on the model, such as the timing of redescription and domain-specific differences.

Chapter 3

Connectionism and Developmental Modelling

3.1 Introduction

This chapter surveys computational models of development, comparing connectionist models, the focus, with symbolic and dynamical systems approaches. The second half of the chapter reviews requirements and previous suggestions for a computational model of the RRH, discussing related connectionist issues, in particular systematicity, explicitness and task transfer, which such an enterprise raises. Practical investigations into modelling the RRH using resource-phased connectionist models are reported in chapters 5–7.

3.2 Computational models of development

As discussed in the closing sections of chapter 2, computational modelling has been advocated for developmental study for several central reasons. Klahr (1995) notes two clarifying roles. Firstly, a given developmental theory may be 'sufficiently complex that only a computational model will enable one to derive predictions from it' (p. 358), and secondly, in comparison to verbally expressed theories in particular, such models afford testable, explicit formulations whose assumptions are evident, and may thus be compared against each other in detail (although it could be argued that certain aspects of connectionist models, for instance the opaque nature of the products of learning, still lead to problems of interpretation, albeit in a different way to those arising from verbal theories such as Piaget's).

The 1990s has seen the main rise to prominence of both developmental modelling and computational approaches to developmental psychology (e.g., Rutkowska (1993), Karmiloff-Smith (1992b)). For instance, Boden (1988, p. 213) notes the comparative rarity of both these enterprises at the time of writing. In his chronology of computational models of development, Klahr (1995) notes that despite the appearance in the early 1970s of production-systems models of performance at distinct developmental stages, process models (here in the form of self-modifying production systems such as Wallace, Klahr, and Bluff (1987)) did not appear until the early 1980s.

Developmental models may be divided along essentially similar lines to computational models of cognition in general. Klahr (1995) distinguishes between production-system, connectionist, and *ad hoc* models (i.e., those intended to capture the fine-grain behaviour in a particular domain or task without commitment to a particular cognitive architecture or modelling paradigm). The following sections review and compare these approaches as well as models conceived of in the dynamical systems paradigm.

3.2.1 Symbolic Models

Although many symbolic models are based on production systems architectures (discussed below), other styles of symbolic model also exist. Klahr (1995) discusses what he calls *ad hoc* models, in the sense that they are not constructed within a specific framework such as a production system, although they may give a finer-grained fit to the data, for instance the Structure-Mapping Engine of Gentner et al. (1995). Schmidt and Ling investigate the use of traditional machine learning techniques such as discrimination trees for modelling development in the domain of the acquisition of the English past tense (Ling, 1994) and the balance-scale task (Schmidt & Ling, forthcoming).

Discovery learning models

Although not explicitly intended as developmental models, discovery learning models are a class of symbolic models with particular relevance to some of the ideas of the RRH. Examples include BACON (Langley, Bradshaw, & Simon, 1983), a data-driven model of scientific discovery; COPYCAT (Mitchell & Hofstadter, 1990), which applied analogical mappings in a domain of simple letter strings; AM (Lenat, 1982), and EURISKO (Lenat, 1983), which investigated theory-driven discovery. The AM program modelled discovery of concepts in the domain of number theory, while its successor EURISKO was an attempt (which was less successful) to capture concepts about heuristics themselves. Of these models, BACON, AM and EURISKO may be related to the ideas about the role of theorising in learning put forward by Karmiloff-Smith (1988) and Karmiloff-Smith and Inhelder (1975). EURISKO also represents an attempt to capture the reappropriation of knowledge associated with metacognition.

Production Systems Models

A *production system* is a computer model stated in terms of condition–action rules. The basic structure of a production system comprises two memory spaces:

- Working memory — a collection of symbol structures called working memory elements
- Production memory — condition–action rules (or productions) which consist of conditions corresponding to patterns of working-memory elements, and actions which modify the contents of working memory

This structure is used in conjunction with a cyclical condition-matching and action-execution process, and a mechanism which resolves conflicts between any productions whose conditions have been satisfied simultaneously. The handling of conditions provides a parallel associative recognition memory while the actions take place serially.

The main advantages of production systems are outlined by Neches, Langley, and Klahr (1987, p. 13). These include the fact that rules tend to carry equivalent amounts of information, the psychologically realistic combination of serial and parallel processing inherent in the matching and execution procedures and the ability to model long- and short-term memory and the relationship between them. Of particular relevance to developmental modelling, according to Neches et al., is the independence of rules, as this facilitates the addition or removal of rules and hence the system's ability to capture incremental change and successive developmental stages.

The general strengths of the production systems approach include its ability to capture fine-grained behaviours and its ability to model strategy change (see below).

Although earlier models were used to capture only static performance, learning and development were subsequently introduced (Neches et al., 1987). Production systems have been used to model the development of cognitive competence in domains such as children's counting and mathematical skills (Wallace et al., 1987), Piagetian tasks such as seriation (Young, 1976) and transitive reasoning and grammar acquisition (Langley, 1987). More sophisticated systems also make use of augmentations such as production weightings and self-modifying productions (Wallace et al., 1987).

Production system models of developmental change include Langley's discriminant learning model of stage-transitions on the balance-scale task (Langley, 1987), and Wallace et al.'s self-modifying production model of children's number sense (Wallace et al., 1987).

Some workers in this field (e.g., Anderson (1983), Newell (1988)) also make strong claims that production systems correspond to the *cognitive architecture* (defined by Neches et al. (1987, p. 14) as 'the invariant features of the human information processing system') underlying human cognition.

3.2.2 Dynamical Models

Several models intended to capture Piagetian stage phenomena have also been constructed in dynamical systems terms. The models proposed by Preece (1980) and van der Maas and Molenaar (1992) are based on the notion that qualitative changes in catastrophe theory provide a basis for reasoning about qualitative changes during development, in the absence of any discussion of representation, but in a more abstract manner than the dynamical systems framework of Thelen and Smith (1994) discussed in Chapter 2.

3.2.3 Connectionist models

Although connectionist models are discussed in more detail in sections 3.3–3.6, it is worth surveying here the qualities which, it is argued, make them appropriate for modelling development.

Karmiloff-Smith (1992a, p. 4) emphasises the qualities of connectionist approaches which have particular relevance to her work; specifically their potential as a means to analyse implicit representations, since connectionist models do not rely on the explicit codings often underlying performance in traditional cognitive models. Like Mareschal and Shultz (1993), she also points to the gradualism and non-linearity of connectionist models and the way this changes ideas about stage transitions, as well as allowing systems to avoid premature commitment to hypotheses. As discussed in section 3.4.1, Karmiloff-Smith also sees networks as implementing a kind of progressive modularisation in the form of increasing informational encapsulation.

Such models also take advantage of some of the inherent qualities of connectionism considered relevant to models of cognition in general. For instance the fact that networks simultaneously learn by rote and extract graded generalisations and that the representations they develop are graded and distributed, exhibiting graceful degradation and saturation.

3.2.4 Discussion

Comparing production systems with connectionist models

Cognitive architecture As Klahr (1995) points out, implicit in production systems models is a strong claim about cognitive architecture, while connectionist models, according to Klahr 'are less of an architecture than a set of shared assumptions' Klahr (1995, p. 363). In comparing the two approaches, he goes on to argue that properties such as parallelism and distribution of representations, usually claimed as advantages for connectionism, are also inherent or possible in production systems.

Capturing change Boden (1988) in reviewing computationally inspired answers to the question of the difference in abruptness supposed to exist between learning and developmental change, notes that adding a single rule to a production system model can lead to a qualitative change in behaviour 'comparable to what Piaget would term a stage progression' (p. 211). It should perhaps be remembered here that productions vary greatly in the granularity and abstraction of knowledge they embody. Thus a single production rule may well capture a crucial strategy change in itself, in a way in which a connectionist training pass in particular typically does not, except perhaps in special cases where learning is one-shot or semantically transparent (Clark, 1989). However, Klahr (1995) argues against the intuition that a change in the rule-base of a production system must always be viewed as a qualitative change at a much higher level of

granularity than a change to some component of a connectionist system. He claims instead that the granularity of an individual production rule may be sub-symbolic or implicit, and that, using productions with continuously varying strengths, gradualistic behaviour similar to that usually exhibited by networks, which as he points out, can also exhibit dramatic change.

Critics of the traditional (symbolic) approach to cognitive modelling (e.g., Bates and Elman (1992), Plunkett and Sinha (1992)) tend to regard it as antidevelopmental in more fundamental ways. For Plunkett and Sinha, the three notions of an essentially private, internal and representational cognition, functionalism and what semantic transparency, together with a perspective on development which is essentially ahistorical, make the cognitivist programme the antithesis of the epigeneticism they argue for.

Bechtel and Abrahamsen (1991), although regarding connectionism favourably as a means of modelling development, take a less radical view of the divisions between traditional and connectionist approaches. They point out that some of the same shortcomings of traditional models which helped to prompt the renewed interest in connectionism in the early 1980s (e.g., poor generalisation, inflexibility, and particularly brittleness) also provoked a new wave of symbolic models, which Bechtel and Abrahamsen refer to as 'non-traditional' symbolic models. The modifications in approach which characterise such models include the use of finer grain 'microrules', parallel rule-selection and/or activation, soft constraints — effected by adding strength parameters to rules and using these as part of the criteria for rule-selection, resilience via redundancy, greater attention to learning algorithms, knowledge compilation, chunking and rule transfer.

Wallace et al. (1987)'s BAIRN model is an example of a non-traditional symbolic developmental model, claiming to provide an integrated model of learning and development, using a network of knowledge-communicating nodes which undergoes self-modification as a result of its interactions with the environment, i.e., the input-data with which it is presented.

Although differences such as the use of explicitly sequenced symbol strings and operations, and non-local control still exist between connectionist and non-traditional symbolic models, the revised approach has served to bring the two programmes closer. Bechtel and Abrahamsen (1991) regard these differences as being small enough that 'empirical adequacy will not be the primary determinant of the fate of symbolic versus connectionist models' (p. 18).

Implicit representation As Karmiloff-Smith (1992a) points out, it seems more appropriate to investigate implicit learning on the basis of representations which are not explicitly coded in the way in which production rules tend to be.

Production system models are at a higher granularity than connectionist models, are inherently not semantically opaque, and provide no way of explaining the sub-symbolic–symbolic transition.

Production systems as models of RR Boden (1988) points out that ACT* allows rules to be fired as well as matched in parallel, resulting in a redundant multiplicity of representations (as the RRH requires). However ACT* is essentially a model of knowledge compilation and the progressive automatisation of processes which Karmiloff-Smith (1992b, p. 17) contrasts with the RRH.

Some production system models exist, such as the HPM (Heuristic Procedure Modification) framework (Neches, 1987) which seem closer to ideas in the RRH. HPM emphasises the role of invention in learning and the redeployment of the components of past learning, contrasting the latter with proceduralisation. However, although such a model might be able to capture the transitions between the levels of explicit representation in the RR model, again it could provide no explanation of the initial implicit–explicit progression.

Process models

A criticism which could be directed at early production system models of developmental phenomena, for instance, is that, as Neches et al. (1987, p. 18) say, 'these stage models ... explained behavioral differences at successive stages in terms of slightly different rule sets, but yet provided

no *mechanisms* to account for the transition process.', and as mentioned above, advocates of the epigenetic approach such as Plunkett and Sinha are critical of symbolic approaches partly for the synchronic modelling they seem to encourage. Although connectionist models are seen as closer to the ideal of true process models of development, we must perhaps still be cautious about ascribing developmental properties to learning systems, or as Clark (1993a) puts it, to differentiate between learning trajectories and developmental trajectories.

The following sections present several examples of connectionist process models of development in more detail, focusing on qualitative change, occurring both macro- and microdevelopmentally.

3.3 Connectionist models of qualitative change

3.3.1 Modelling stages

There have been a number of attempts to use connectionist models to capture aspects of Piagetian stages by trying to replicate the results of various experiments concerning tasks such as seriation (Mareschal & Shultz, 1993) and torque-difference (McClelland, 1989; Shultz & Schmidt, 1991).

McClelland (1989) used a modified backpropagation architecture in which the learning of the torque-difference effect (in this context the rule that a balance-scale will balance if the product of the weight and distance from the pivot is equal on both sides) is modelled by presenting a localist encoding of the weight on each side of the balance as well as its distance from the pivot, in conjunction with training information about whether the scale goes down on the left, right or balances. The standard backpropagation learning rule is used to train the network which successfully learns to use the comparison of the two weight–distance products as a means of determining whether the scale will balance, even in cases where distance and weight both differ on each side. The model also exhibits a stage-like progression through different strategies, first using only information about weight alone, then distance alone and finally using the full torque-difference rule.

The model includes several assumptions including the modification of the architecture so that the inputs relating to weight are connected to a distinct set of hidden units from those representing distance although the network does learn to partition these weight and distance units according to which units correspond to which side of the balance). According to McClelland, these structural constraints were necessary in order for the network to exhibit the desired stage progression.

Some biasing is also inherent in the presentation of the training data, in that early on more examples of problems in which only the weight differed while distance was kept constant) were presented, followed by a phase in which equal-weight/differing-distance problems predominated, moving finally to a stage where the network is exposed to the most difficult problems in which both values differ, i.e., those which the concept of torque-difference is necessary to solve.

Considering the careful manipulation of training schedule along with the explicit hand-structuring of the network, it is somewhat unsurprising that the network focuses first on the equal-distance problems which predominate early in the training data and transfers its focus of attention to equal-weight problems. There are no domain-specific constraints built into the network which make it attend to the visual modality first — this is built into the training data. Shultz and Schmidt (1991) attempted to obtain the same results without recourse to the same weight of assumptions. Their study also made use of training-data manipulation, using a data set that was both gradually expanded over the course of training as well as including a bias towards equal-distance problems. An important contrast between their work and McClelland's is that they used the cascade-correlation architecture (see 4.1) which is inherently capable of shifts in representational power.

The training bias was effected by making the random selection of the 100 initial training

patterns (of a total of 625 possible configurations) subject to a bias of 0.9 in favour of selecting an equal-distance problem. At each output epoch the training set was expanded by one pattern drawn at random with replacement and subject to the same bias towards equal-distance problems. Shultz and Schmidt note that these measures were found to be necessary if the network is not to find learning extremely difficult and also, as in the McClelland (1989) model, if the model is to exhibit any behaviour other than the final torque-difference strategy.

Mareschal and Shultz (1993) again used cascade-correlation in their model of the seriation task, proposed by Inhelder and Piaget (1969) as a test of transitive reasoning involving the sorting of elements according to their their relative values on some scale or dimension, e.g., different-length sticks according to length.

The stage phenomena to be captured in this task are as follows. At the (first stage circa age 4) children make no real effort at ordering, making only random movements of the sticks. By age 5 they are able to perform localised orderings of two or three sticks at a time according to absolute quantities whether a stick is 'big' or 'small'). The third stage (circa age 6) involves the ability to construct a series with difficulty through trial and error, while at the final stage children are able to use a strategy to perform the task efficiently.

The results of this study conformed to the above pattern, with stages often overlapping in time. Of 20 network trials run, 7 exhibited all four stages, with the majority of the remainder exhibiting either 1, 2 and 4 or 1, 3 and 4. The model was also shown to respond to perceptual variation in a plausible way with stage-3 performance becoming stage-4 when the relative difference in length between the sticks was increased.

Stage differences were not found to be marked by large differences in the weights, although the network was observed to adjust weights relating to the short end of the series first, progressing to the larger end, an effect which Mareschal and Shultz see as consistent with findings that children build linear-order representations from the ends inwards.

Although previous computational models of the seriation task exist, e.g., Young's (1976) production-system model, in which the development of an individual's performance is modelled by a continuous process of new-rule acquisition according to the selection and evaluation of items and the correction of incorrect choices, the authors claim that

> [n]one of these models are truly developmental since they do not provide a mechanism for passing from one stage to the next.
>
> (Mareschal & Shultz, 1993, p. 2)

Shultz (1991) argues in the context of the cascade-correlation models presented above that connectionist models which exhibit stage behaviour tend to do so in a way which exhibits several desirable properties which are less characteristic of rule-based models — transition to a higher-level stage is typically soft and tentative, there is some stage-skipping and a limited amount of regression to earlier stages. Network models are also comparatively successful in capturing the ordering and organisation of stages.

Causes of stagelike transitions in networks

Shultz et al. (1995) attribute the ability of cascade-correlation to capture stagelike behaviour in the seriation, balance-scale, personal pronoun, and time-distance-velocity tasks to a variety of factors, in particular a combination of initial training-data biases and the changes in representational power due to the recruitment of a small number of hidden units. The additional bias due to the modularisation of selecting and moving tasks also contributed to the staged behaviour in the seriation task.

McClelland (1995) also attributes the stagelike progressions observed in backpropagation networks to a combination of initial bias and changes in learning speed observed in standard backpropagation networks. For instance in the case of the balance-scale model:

> qualitative changes — apparent stagelike progression — could arise from the accumulation of small incremental changes. In the model, acquisition of the use of each of two cues [weight and distance] begins with an initial phase in which the effects of experience accumulate gradually, followed by a more rapid acceleration.
>
> (McClelland, 1995, p. 193)

Although not aiming to model stages as such, Plunkett and Marchman (1993) also note that in the context of learning the past tense of English verbs, improvements in performance accelerate rapidly after an initial period of gradual change. The frequencies of regular and irregular verb form were also central to the pattern of qualitative change (consisting in learning, overregularisation and eventual correction) observed.

Shultz (1991) highlights several features which contribute towards the emergence of stagelike features in networks, including hidden-unit herding (Fahlman & Lebiere, 1990), overgeneralisation, and the hidden-unit recruitment of generative architectures, all of which are seen as ways in which networks form partial solutions to problems. This last is central in Shultz's view to network models of qualitative stage-transitions. The generative capacities and attendant increases in representational power of networks such as cascade-correlation are implicated in the ability to reach and stay in stage 4 of the balance-scale task, requiring a grasp of the principle of torque (other models were unable to reach this stage or were able to only at the expense of earlier stages).

Qualitative change in networks and Piagetian stages

Several of these connectionist workers attempt to relate the qualitative change in their models to that which occurs in Piagetian stage-transitions, considering connectionism as a potential mechanistic explanation (although Klahr (1995) is critical of this enterprise on the basis that Piaget's original formulations are not well specified enough to permit such a comparison to be made on a scientific basis). For instance, McClelland (1995, p. 193) compares the accumulation of incremental change to give (macro-developmental) qualitative change in connectionist networks to Piaget's (verbal) description of equilibration.

Shultz et al. (1995, p. 255) claim that learning in cascade-correlation can be described within a Piagetian framework in terms of assimilation, assimilative learning and accommodation. Pure assimilation corresponds to correct generalisation to new exemplars without further learning or structural change, and assimilative learning to learning which modifies weights but not structure. Accommodation occurs when the network is forced to increase its representational power through recruitment of a new hidden unit.

Shultz (1991) notes important differences between Piagetian stages and qualitative transitions in networks, in particular that network changes are gradual (rather than abrupt), task- or domain-specific rather than broad-based and domain-general, and timed according to task, rather than occurring concurrently across domains (although it is difficult to think of a connectionist model which investigates development across unrelated domains; despite the existence of some models relating the concurrent development of the verbal and visual or verbal and conceptual domains (e.g., Schyns (1991)), none seem to exhibit stages. Despite these differences, Shultz notes that networks are capable of capturing other aspects of stages including invariant ordering and organisation.

3.3.2 Modelling U-shaped behavioural curves

Modelling the acquisition of the English past tense

A model which has become one of the most controversial in discussions of connectionist learning and representation is that of the learning of the English past tense by children originally constructed by Rumelhart and McClelland (1986). This model was an attempt to account for the way in which English-speaking children tend to produce errors in their formation of the past

tense of verbs such as 'sitted'. The model also sought to reflect the observation that the learning of this domain is often observed to produce a U-shaped behavioural curve (Strauss & Stavy, 1982).

The model uses the perceptron learning rule to form associations between a layer representing the verb stem and another layer representing the past-tense form, both of which are encoded using a pronunciation-based scheme.

The model's apparent ability to capture many of the overregularisation and U-shaped performance effects found in the experimental literature, for instance the way irregular verbs are treated as regular verbs during the early stages of learning — an effect resulting in the 'trough' of a U-shaped curve, seems impressive, and its creators considered that the models ability to exhibit these phenomena without recourse to rules and without a separate mechanism being necessary for treating irregular and regular verbs could be, as Plunkett and Sinha (1992) put it:

> interpreted as challenging the traditional view that acquisition is necessarily a process of organizing and reorganizing explicitly represented rules, with a separate representation of exceptions to the rules.
>
> (Plunkett & Sinha, 1992, p. 224)

In the now-familiar critique by Pinker and Prince (1988) the main criticisms of this model are that there are fundamental dissimilarities between the task which the system has to learn and that faced by children, who are not exposed to stem and past-tense forms occurring side-by-side in the input in the absence of semantic information, for instance. It is also argued that the U-shaped curves are not caused by anything more than the manipulation of the size of the data set. In particular, training begins with presentation of one instance of each of the ten most common English verbs, eight of which are irregular and this initial training phase is followed by both an increase in the number of regular verbs as less common verbs are included in the training corpus as well as an increase in the overall size of the training corpus. It is argued that the U-shaped curve is a direct result of this change in the training data and that it is unsurprising that a degradation in performance on irregular verbs should coincide with the point at which the network begins to see many more examples of regular verbs.

More recently Marchman has responded to these criticisms by constructing network models of the same phenomena which do not rely upon the same assumptions regarding the manipulation of training-set size. For instance, Plunkett and Marchman (1989, 1991) showed that U-shaped patterns were observed even if the training corpus remained fixed. This is attributed to the fact that conflicting mapping types exist in the training corpus and that it is the network's attempts to resolve these conflicts that lead to the temporary degradations in performance observed on particular categories. This would seem to be explained by what Fahlman and Lebiere (1990) call 'hidden-unit herding' or the 'moving target' problem (see section 4.1).

Plunkett and Marchman (1989) also discovered evidence that U-shaped patterns are neither restricted to a single period of development nor constrained to occur simultaneously for different categories, or indeed not to occur multiple times even in the case of a single verb. All these findings suggest that U-shaped patterns might be best considered a micro- rather than a macro-developmental phenomenon.

Incrementing the training corpus one verb at a time yielded improved performance in the mapping of verb-stems to past-tense forms. Plunkett and Sinha (1992) also note two other related points which are important in this context, firstly that changes in the way verbs are represented occur when the corpus reaches a certain 'critical mass', and secondly that these changes result from an internal pressure towards a generalisation of the early form of the network. They note that performance on the newly added verbs early in training is difficult to classify, adding that

> [t]his result demonstrates that the form of representation underlying the network's successful mapping of the initial set of 20 stem/past-tense pairs does not generalize

> well to new forms and that these initial 20 verbs are essentially memorized by the network by a process we can refer to as rote learning.
>
> (Plunkett & Sinha, 1992, p. 227)

and concluding that

> later in training, the network's representations become systematized (as evidenced by the performance on novel verbs) ... the network continues to map irregular verbs correctly even though the mapping of novel verbs is systematic.

These results support the important claim that learning and generalisation can be realised within a single mechanism.

It could be argued however that a network implementing a rule-plus-exceptions scheme should no longer be regarded as utilising a single mechanism. Indeed in some cases, the solution formed by a network may be a very close approximation to an explicit mechanism in terms of its classification behaviour. But even if we reject the claims made by Pinker and Prince (1988) that connectionist systems never fully implement the equivalent of categorical symbolic rules or even rote-memorisation of exemplars, it should be noted that, whatever their behaviour when fully trained, the kind of connectionist systems under discussion here begin with a single, undifferentiated mapping strategy, and only come to realise the two kinds of behaviour through learning. By contrast, a symbolic system of the kind outlined by Pinker and Prince (1988) is equipped from the outset with separate and explicit mechanisms for rote-memorisation and rule-acquisition (via the generation and testing of hypotheses). In terms of developmental modelling, it seems we stand to learn more from the emergence of full or partial approximations to classical systems than by assuming their presence, although the issue can also be seen as one of the extent to which symbolic mechanisms are innate.

3.4 Other issues for connectionist developmental modelling

In the sections which follow I discuss a number of (somewhat interrelated) issues which are important in current work on connectionist developmental modelling. Also particularly pertinent to the enterprise of constructing a model of the RRH are the issues of how to capture the increasing explicitness and accessibility of knowledge as well as the different sources of that knowledge (see 2.1).

3.4.1 Innateness

Connectionism has often been criticised as implying an empiricist standpoint, with modellers regarding their efforts as existence proofs that learning can proceed from an initial *tabula rasa* (Bechtel & Abrahamsen, 1991). Karmiloff-Smith (1992a, p. 23) comments that although some connectionists would deny that such a viewpoint is inherent in their research, their approach is still comparable in some sense with a Behaviourist position, in that it allows for innate structure while ignoring any need to specify innate content.

Plunkett and Sinha (1992, p. 250) argue that a connectionist position 'is in no way equivalent to a *tabula rasa* account', but is rather aimed at establishing the minimum necessary initial conditions for the emergence of rich behavioural and representational properties in relation to a particular environment.

Connectionism and domain-specific constraints

In her discussion of this topic, Karmiloff-Smith (1992b, p. 181) notes that in being applied only to single tasks, networks are inherently domain-specific and that, although they are informationally encapsulated, their learning echoes progressive modularisation rather than prescribed Fodorean modularity.

However she also complains (Karmiloff-Smith, 1992b, p. 188) that connectionist models are in fact task specific (or in other words *micro*domain-specific). In more recent discussions (Karmiloff-Smith, 1994) she has softened this claim, agreeing with Shultz (1994) that modelling would do best to proceed by trying to integrate several tasks before aiming for some abstracted notion of 'task-independent' learning.

But apart from restricting exposure to data from particular domains, it is difficult initially to see how constraints which are truly domain-specific with respect to content but precede learning are to be incorporated into connectionist nets.

Karmiloff-Smith (1992a) suggests that training-set design and the biasing of initial weights to simulate innate attentional biases constitute examples of this, citing as an example of the latter the initial bias towards weight information in McClelland (1989)'s simulation of children's performance on the balance-scale.

Another way to simulate such biases within a purely developmental model may be by pre-training and freezing a preliminary recoding subnet. More naturalistically, and truer to the idea of innate (rather than early-acquired) constraints, evolutionary techniques could be used to produce a network with the appropriate initial biases (see Nolfi, Parisi, and Elman (1994) for example).

Karmiloff-Smith (1992a) also distinguishes between representational and architectural biases to development in networks. A network such as that of Elman (1991), she claims, incorporates both architectural biases in terms of the choice of a locally recurrent network model as well as some representational biases in the form of the meta-linguistic assumptions made by Elman in designing the training set.

In methodological terms, Karmiloff-Smith (1992a, p. 23) speculates that (despite the orthogonality of the domain-general–domain-specific issue to the underpinnings of the connectionist framework) the fact that, in her view, 'in practice the notion of domain-general learning algorithms and knowledge-free starting points ... has been championed by connectionist modellers'(p. 31). may be due to the often adverse reactions of nativists (e.g., Pinker and Prince (1988)) to connectionist work. I suggest further that the desire to move away from the hand-crafted nature of traditional symbolic models may also contribute to the strong drive to avoid building initial content into connectionist models.

Connectionism and domain-general constraints

Domain-general constraints seem easier to relate to connectionist models. Karmiloff-Smith (1992b, 1992c) sees in connectionist work on developmental theory a claim to be producing domain-general models, on the basis that single architectures are put forward as capturing behaviour or development in several domains (see Shultz et al. (1995) for instance). Karmiloff-Smith (1992c, p. 257) points out that despite using the same general class of learning algorithms, in practice any given model is distinguished by architecture, choice of learning rules, initial parameters, learning rate and so on.

The following sections deal with problems which have been identified with connectionist cognitive modelling in general (Fodor & Pylyshyn, 1988; Clark & Karmiloff-Smith, 1993; Halford, 1993), but are of particular relevance to the RRH as well developmental modelling.

3.4.2 Systematicity

The issue of whether (and secondarily to what extent, and how inherently), connectionism can capture the systematicity of knowledge is one which now pervades connectionist study. Systematicity was highlighted in Fodor and Pylyshyn (1988)'s critique as a central problematic capacity for connectionist cognitive modelling. The relationship between development and systematicity is discussed in section 3.5.1 below.

There are several responses to this challenge. Clark (1989) makes the important observation that systematic behaviour need not imply systematic, and specifically classical symbolic, innards. It is also questionable whether the full-blown systematicity which Fodor and Pylyshyn assume

is as pervasive as they claim. For instance in the RR literature we see that (for reasons yet to be determined) not every domain is redescribed into systematic form, while in other domains, such as adult language learning in particular, modularisation prevents redescription. Bechtel and Abrahamsen (1991, p. 235) note that even in (nominally systematic) adult sentence-processing asymmetries remain which reflect developmental differences.

A third response, related to Clark's above, is to concentrate on the kinds of systematicity which connectionist systems capture naturally. As we saw above, Plunkett and Marchman (1993) found that backpropagation networks were forced to develop (partially) systematic hidden representations when the number of training items increased beyond that which the network could represent by 'rote' (i.e., as a form of look-up table). This kind of systematicity would seem to correspond to Kirsh (1991)'s intermediate level concepts, which may be used for prediction (in the model, the correct generalisation to novel forms) but are not accessible to consciousness, and thus also perhaps to level E1 in the RR model.

Clark (1993a) argues that this typically connectionist kind of partial systematicity can be viewed as a legitimate intermediate representation on the way to a state which approximates full syntactic systematicity in the way human adult performance seems to demand, but differs fundamentally from that conceived of by Fodor and Pylyshyn in that it is the product of, rather than the prerequisite for, a developmental process.

3.4.3 Transfer of learning

In chapter 2 we saw that one way of characterising the later levels of redescription was according to their *accessibility* to the processes associated with other tasks or domains. Transfer of learning between networks is an obvious way of operationalising this abstract idea of accessibility.

For the purposes of this discussion, studies of transferability in networks can be divided into two classes: those in which problems are clustered according to their similarity with respect to a certain learning scheme (e.g., Pratt (1993, 1994), Sharkey and Sharkey (1993), Thrun and O'Sullivan (1995)), and those which attempt to transfer learning to tasks which are structurally related (e.g., Dienes et al. (1995), Dienes, Altmann, and Gao (submitted)). Although both of these kinds of methods imply a space of related tasks, the ways in which they are related is fundamentally different, corresponding in some ways to the associative–relational divide identified by Clark and Thornton (1993), Philips et al. (submitted) in that in the first class, similarity maps onto actual proximity in the hyperspace defined by the network, while in the latter similar solutions are likely to be distributed throughout the space.

3.4.4 Explicitness

Although the semantic opacity of standard connectionist systems is well-acknowledged (Clark, 1989; Karmiloff-Smith, 1992b), the extent to which networks can represent explicit knowledge is still in question.

Kirsh (1991) distinguishes between three levels of conceptual knowledge: the first is used for recognition of perceptual features, the intermediate level allows prediction but not conscious analysis, while the third is what might more usually be referred to as conceptual knowledge, and facilitates full-blown compositionality, systematicity and expressibility.

I would argue that, for connectionist systems, explicitness, like accessibility and systematicity, is both continuous and relative to the particular system(s).

For instance Plunkett (1993, pp. 554–5) points out that representations implicit in one network may be explicit to another and that although the representations may be opaque and implicit to the experimenter, the network reacts directly to them. Thus, he emphasises that 'one must evaluate whether the network has constructed an implicit or explicit representation *in the context of the task to be performed*. (p. 555). Karmiloff-Smith and Clark (1993) concur with this continuous view, relating it to the representational multiplicity inherent in the RRH. This idea

of explicitness leads Clark (1993a) to equate (or operationalise) explicitness as some function of accessibility and multiple usability.

3.5 Incremental learning

Incremental learning (Bates & Elman, 1992; Elman, 1991) is a broad term used to refer to a collection of (mainly connectionist) methods in which (in the most general terms) learning is staged in some way such that the complexity of the overall task to be learnt is reduced. These techniques have arisen in both engineering and cognitive modelling settings primarily as ways of facilitating learning in otherwise intractable scenarios (see Elman (1991) for instance), and also represent the major technique (besides the choice of basic learning algorithm) used in the connectionist simulation of development (Shultz et al., 1995; Elman, 1991).

Examples of incremental methods include manipulations to resources such as training set or units (e.g., cascade-correlation models (Shultz et al., 1995)), connections, or subnetworks, or manipulations to training corpora, typically increments (Plunkett & Marchman, 1993, 1991; Rumelhart & McClelland, 1986; Shultz & Schmidt, 1991) or changes in composition by token type (Rumelhart & McClelland, 1986). Some of these, e.g., the phasing of training-set difficulty and incremental increase in attention in Elman (1991) are considered to be functionally equivalent, but whether this applies to all the above methods in more than the most general terms is, as yet, an open question.

In terms of biological realism, architectural resource phasing can be seen as a loose analogue of maturation. Progressive pruning of components of a network architecture may also correspond at some level of abstraction to selectionist processes which accompany brain maturation (see e.g., Johnson and Karmiloff-Smith (1992)).

More recently Clark and Thornton (1993) have refined the notion of incremental learning, introducing a distinction between what they term *conservative* and *extended* incremental solutions. A body of training data constitutes a conservatively decomposable problem if it is possible to reduce the search space for subsequent learning by focusing at first on a subset of that same body of data, while if the lower-level feature-detectors (or other recodings necessary for learning later complexities) must instead be developed through attempts to perform some other task (i.e., in cases where this cannot be achieved by any amount of exposure to the original data), then the problem has only what Clark and Thornton (1993) call an extended incremental solution.

Both of the resource-phasing techniques presented in Elman (1991) constitute examples of conservative incremental methods. Batching and grading training examples and placing initial limitations on the network's attentional window both involve presenting the network initially with a simpler subset of the original training data in way which fortuitously provides a basis for learning the data in its full complexity. Spatial modularisation schemes such as that of Jacobs, Jordan, and Barto (1991) are cited by Clark and Thornton (1993) as examples of extended incremental learning methods.

3.5.1 Developmental trajectory

Related to the idea of incremental learning is that of a developmental (or representational) *trajectory* (Clark, 1993b, 1993a). Fundamentally the idea is that, in order to learn complex or hierarchically structured problems, a learning system must pass through an ordered set of configurations of increasing power (termed by Clark (1993a, p. 151) the cascade of significant virtual machines).

Trajectory and Systematicity

Clark (1993a, p. 149) argues persuasively against Fodor and Pylyshyn (1988)'s concept of systematicity, which he re-expresses as

> a notion of closure of a set of potential thoughts under processes of logical combination and recombination of their component "parts"

(Clark, 1993a, p. 147)

He proposes that instead of viewing systematicity as something built into an underlying cognitive architecture, we treat the space of interanimated concepts as another complex space and systematicity as a knowledge-driven concept which the network must acquire through learning:

> The mature knowledge of such a system will be expressible in terms of a (largely) systematically interwoven set of concepts. But the systematicity will be learned as a feature of the meanings involved. It will not flow from the shallow closure of a logical system under recombinative rules, but from hard-won knowledge of the nature of the domain.
>
> (Clark, 1993a, p. 149)

Clark goes on to suggest that connectionist systems need to be scaffolded in order to learn about complex spaces. The term scaffolding refers here to the process of supporting the progress of a learner through a suitable series of configurations, for instance by incremental learning. This implies also that systematicity (in the acquired sense described above) might require scaffolding.

3.6 Hybrid models

The term *hybrid* refers primarily to connectionist models which make use of a component from a different modelling paradigm in order to take advantage of a different style of processing. The most common form of hybrid models (in this sense) are connectionist–symbolic hybrids which use a traditional symbolic component in order to provide systematicity or increase explanatory power, e.g., in engineering contexts such as expert systems to introduce or identify human-recognisable rules after training (e.g., Craven and Shavlik (1994)).

Another sense of hybrid refers to connectionist schemes which make use of a mixture of representational styles (e.g., localist and distributed), learning strategies or architectures (e.g., Schyns (1991)), or consist of a number of modules between which data is communicated in a way which is not essentially connectionist. In this weaker sense, many modern connectionist schemes such as cascade-correlation (Fahlman & Lebiere, 1990) may be considered hybrid. It is likely that it is these models to which Karmiloff-Smith (1992b) refers when she speculates that in time the term 'hybrid' will lose force.

Plunkett and Sinha (1992, p. 251) argue against hybrid models (in the connectionist–symbolic sense) that although they seem to provide both connectionist symbol-grounding and the full-blown systematicity of mature human cognition, truly symbolic systems (those which exhibit 'functional univocity') cannot derive from anything but other such systems, and this leads to the same kind of strong nativism required by purely cognitivist accounts, which in turn goes against their vision of an epigenetic connectionism. Another objection to these hybrids is that they can provide no explanation for the emergence of symbolic from sub-symbolic cognition since a mechanism analogous to an external symbolic system is already assumed. It also seems that there is no consensus about how the symbolic and connectionist components should be related, making comparison between systems difficult.

3.7 Specific requirements for a model of RR

Although the RRH is put forward as 'a framework — rather than a precise theory — for exploring possible generalities in developmental change across a range of domains.' (Karmiloff-Smith, 1994), and makes no detailed commitment to any possible mechanisms for redescription itself, certain general requirements have been set out in the literature. For instance in Clark and Karmiloff-Smith (1993, p. 509), the following are presented as requirements for a developmental model (specifically a connectionist one, although they consider others) in the spirit of the RRH:

- the model should treat its own representations as objects of manipulation
- do so independently of prompting by continued training inputs
- retain copies of the original networks
- form new structured representations of its own knowledge which can be manipulated, recombined and accessed by other computational processes

Thus what is emphasised here is that any model of RR should be redundant, and the representational change it gives rise to should be endogenously driven, and facilitate access and inspection to previously generated representations both for the original network (learning model) and for other computational processes.

Karmiloff-Smith also makes implementational suggestions which are more specific to connectionism in the context of discussions (Clark & Karmiloff-Smith, 1993; Karmiloff-Smith, 1992b, 1992c), of what current connectionist models lack in terms of developmental modelling. In addition to the lack of tendency or capacity to move beyond their own success, Karmiloff-Smith (1992c) again notes the lack of true early domain-specificity and the relatively small role given to innate constraints.

The shortcomings of what Clark and Karmiloff-Smith (1993) call 'first-order connectionism' (a somewhat imprecise notion intended to cover many common connectionist systems, for instance those whose architectures are not explicitly designed to implement higher-order processing) imply more specific requirements for connectionist models of RR:

- Learning in first-order connectionist systems is purely example-driven and any change reflects the statistics of the input–output mapping.
- Knowledge of rules is always *emergent*, depending on many subsymbolic representations rather than symbolic expressions.
- First-order connectionist systems have no means of analysing their own activity so as to form symbolic representations of their own processing. Their knowledge of rules always remains implicit unless an external theorist intervenes.

These requirements obviously vary in their generality, in particular it is arguable in my view that a specific dictate that networks keep actual copies of previous states or configurations is in keeping with the other suggestions in this context of an exploratory paper.

3.8 Suggested computational models of RR

Since the publication of Karmiloff-Smith (1992b), Clark and Karmiloff-Smith (1993), a number of schemes (whether existing or novel) have been put forward as going some way towards capturing redescriptive effects. In what follows I survey these suggestions, concluding by presenting some previous implementational work which is proposed either as a partial model of the RRH or as implementing similar (but usually more general) mechanisms.
Suggested models of redescription fall into the following broad categories:

- Connectionist — novel schemes i.e., combinations of connectionist techniques not previously devised for another use
- Connectionist — existing schemes
- Connectionist–symbolic hybrids
- Non-connectionist suggestions, e.g., classifier systems

I will discuss each of these in turn.

3.8.1 Novel connectionist schemes

Skeletonisation of copied networks

Clark and Karmiloff-Smith (1993)'s main proposal for a purely connectionist architecture which might be able to capture redescriptive effects involves combining Mozer and Smolensky (1989a)'s *skeletonisation* procedure (hereafter simply 'skeletonisation') with network copying. Skeletonisation involves the pruning of input- or hidden units according to a *relevance* criterion which assesses the importance of particular units in terms of their contribution to the reduction of the overall error — a criterion which is claimed to transcend the statistical profile of the training set.

Skeletonisation has several properties which seem to make it a promising candidate for a purely connectionist model of the RR process. In accordance with the requirements set down by Clark and Karmiloff-Smith (see section 3.7 above), the skeletonisation procedure can be seen as constituting a manipulation of representation, and as acting without the prompting of further training. As Clark and Karmiloff-Smith (1993) acknowledge, the basic skeletonisation procedure would need to be augmented by some form of network-copying scheme in order to make it conservative, (particularly in the precise way that they specify, of course). In terms of the formation of structured representations of the network's knowledge which are then accessible to other processing, Clark and Karmiloff-Smith (1993) cite the claims of Mozer and Smolensky (1989a) for the increased generalisation capacity and simplicity of skeletonised networks, suggesting in turn that by using a skeletonised network as an initial basis for further learning it might be possible to obtain 'a connectionist way of explaining the phenomenon of 'transfer of learning' to a systematically altered but related domain.' (p. 508).

One of the problems with these claims is that the idea of a skeletonised network's being easier to interpret in terms of rules for a human theorist, and that of the suitability of such a network as a way of transferring knowledge can come apart, in that there is nothing 'universal' about the relevances certain input- or hidden units have in the context of their original training in the way that there would need to be for transfer to be facilitated purely by skeletonisation in the absence of the training examples for other domains being presented.

Abrahamsen (1993) makes the similar criticisms of skeletonisation that copying whole nets rarely seems appropriate, partly due to the problems of re-using input representations, and also on the grounds that this does not seem to be the way in which the results of previous learning are copied when children extend their knowledge. This corresponds, she claims, better to a single 'path' through a network, or in other words, a single association or mapping. She also considers that there seem to be no grounds for claiming that skeletonisation serves to (re-)articulate procedural components conflated by the original network.

Both Abrahamsen (1993) and Bechtel (1993) are concerned that skeletonisation cannot be changing the overall scope of the system in a way that Clark and Karmiloff-Smith require, and in a way which transcends first-order connectionist methods:

> All that seems to be happening is that a procedure for pruning and copying is used to create networks which then learn new tasks in the same manner as first-order connectionist networks — by the use of new training inputs. The representations are not being 'operated upon' or 'manipulated' [Clark and Karmiloff-Smith (1993)] (p. 504) in any straightforward sense.
>
> (Bechtel, 1993, p. 534)

and this is a point which Karmiloff-Smith and Clark (1993) seem to concede (p. 574).

Bechtel goes on to argue that in a possible system based on the creation of many skeletonised network copies, the need for a procedure which would identify which of these could be productively used for further learning seems to imply the need for a hybrid system, rather than the extended connectionist systems which Clark and Karmiloff-Smith (1993) envisage. It is difficult to see why such a scheme should be considered any more 'hybrid' than the procedure of skeletonisation itself.

Using competitive learning to extract features from previous learning

An example of the redescription of activations (rather than connection-weightings as in the case of skeletonisation) is the use of a layer of units trained using competitive learning to extract the most salient information from a network previously trained using some error-driven method. This approach was proposed by Bechtel (1993, p. 532) and recent implementational work by Greco and Cangelosi (1996b) (see section 3.9) also fits this general idea. Thornton (1995)'s (hybrid) model also makes use of competitive learning for feature detection in conjunction with an algorithmic component for exploiting relational effects.

This approach corresponds broadly to the redescriptive idea that the products of previous learning are crystallised by a procedure which is not directly driven by the input data and which differs from the original learning procedure. These schemes are also conservative, since in both Thornton and Greco and Cangelosi's versions, connections trained in a previous phase, or using error-driven learning are frozen. In Thornton's scheme, the products of past learning are progressively incorporated into a hierarchical structure. The implementational details of these models are also considered in more detail in section 3.9.

3.8.2 Existing connectionist schemes

Some connectionist workers have reacted to the challenges to first-order connectionism put forward in Clark and Karmiloff-Smith (1993) not by proposing extensions to existing models but by arguing that existing models are not in fact subject to these limitations in the way, and to the extent that Clark and Karmiloff-Smith (1993) suggest.

Backpropagation and emergent qualitative change

In his review of Clark and Karmiloff-Smith (1993), Plunkett (1993) argues that several of the effects which Clark and Karmiloff-Smith (1993) claim would require second-order connectionism already seem to be emerging from the dynamics of processing in a standard backpropagation network. In particular, Plunkett claims that the tendency to go beyond success and the ability to form transferable internal representations are both to be observed in certain backpropagation-based simulations.

The ability to progress beyond success is demonstrated by networks trained to form the past tense of English verbs (see Plunkett and Marchman (1993) for instance). Despite early success, performance in such networks tends to deteriorate with further training, even given an unvarying training set, due to the conflicts in mappings inherent in it. Residual error is another factor which can cause (apparently) successful networks to go through qualitative changes in behaviour during subsequent training.

As evidence for the more controversial claim that internal representations in a backpropagation net can both be incorporated into a network intended to learn a different task, as well as being beneficial to that training, Plunkett (1993, p. 556) describes a network which is first trained to produce the past tense of English verbs by mapping semantic inputs onto phonological outputs, and then trained to produce the plural forms of English nouns using the same input–output coding in the same network. It turns out that the training on verbs transfers positively to performance on nouns. The key to this reusability is of course in part the choice of input and output representations, which, as Plunkett (1993) points out are already explicit, as well as the closeness of the two tasks in the context they define, as he also acknowledges.

Karmiloff-Smith and Clark (1993) respond to these claims cautiously, admitting that it is plausible that effects such as interference may be responsible for causing RR effects in some cases, such as that of the French determiner system where mappings from the form of the indefinite article 'un(e)' to its different functions must be resolved. Clark (1993a, p. 167) is clearer in his disagreement with the idea that the reusability of representations exhibited by backpropagation is of the right kind to model redescription; such reuse neither integrates knowledge nor involves structure-transforming generalisations.

The claim about the effect of residual error raises the important issue of the timing of RR, particularly with respect to behavioural mastery. Karmiloff-Smith and Clark (1993) respond:

> Connectionist simulations have certainly convinced us that full behavioural mastery may not be a prior requisite for the processes of change to start to take place, for indeed one can detect at the hidden layer representations which are not yet apparent at the output layer.
>
> (Karmiloff-Smith & Clark, 1993, p. 573)

The idea that redescription or similar representation reorganisation can take place before behavioural mastery is also supported by the work of Goldin-Meadow and Alibali (1994) and Gentner et al. (1995).

Moving from feedforward to tensor-product networks

The proposal for an RR model presented in Philips et al. (submitted) also makes an alignment between the implicit–explicit distinction in RR and that between associative and relational learning and knowledge representation.

Philips et al. (submitted) provide formal characterisations of associative and relational processing based on specifying data structures and operators. In particular, associative systems consist of data structures containing a set of cue-pairs over a set of primitive symbols, and operators for forming or deleting pairs as well a simple cueing operator. Relational systems consist of a an underlying set of unordered n-'tuples' over the product of n sets of primitive symbols. What distinguishes relational systems is the join, select and project operators. These allow new relations to be formed from all pairwise instances of existing relations, cueing to relate only to a subset of the tuple elements, and individual elements to be accessed on the basis only of their roles, respectively. This latter structure is rich enough to represent recursive data-structures such as trees and graphs and to support the omnidirectional processing necessary for full-blown systematicity.

The implicit–explicit distinction of the RRH is mapped onto this associative–relational distinction. As a step towards unifying the two modes of processing, Philips et al. claim that:

> Since connectionist architectures that exhibit either one of these two modes already exist, connectionism becomes a candidate framework for representational redescription.
>
> (Philips et al., submitted, p. 24)

and propose (in common with Clark and Karmiloff-Smith (1993) and most other commentators) that feedforward networks provide a candidate associative architecture. Tensor product networks are proposed as way of representing relations within a connectionist framework. Briefly, this scheme allows variable bindings and symbol structures to be represented in a distributed manner. The basic idea is that a set of value–attribute pairs can be represented by accumulating activity in a collection of units, each of which computes the product of a feature of a variable and a feature of its value, hence the analogy to the tensor product of two vectors. It is claimed that such networks can be used to represent complex recursive structures, respect the independence of multiple bindings, whilst exhibiting more typically connectionist properties such as graceful saturation. Tensor product networks (Smolensky, 1990) have been used to capture relational or propositional knowledge in Halford, Wilson, Guo, Gayler, Wiles, and Stewart (1994)'s work on structure-mapping during analogical reasoning.

Several criticisms have been made of tensor product representation. McClelland (1995) notes that it is subject to a scaling problem since the number of internal nodes required explodes as the length of the vectors increases. From a classical position Fodor and McLaughlin (1990) complain

both that Smolensky has not provided a truly connectionist alternative to compositionality as an explanation for systematicity, and also that the components of network vectors cannot be involved in causal structure-sensitive processing in the way classical components are, because they remain implicit in explicit instances of the vector as a whole.

Philips et al. (submitted) claim that RR could come about through a process of schema induction in which associations are abstracted to form relations which can then be generalised and applied to structurally similar situations via *alignment* of one structure to the other. But clearly this proposal does not constitute a process model of redescription as it stands since no transition mechanisms are specified. Philips et al. (submitted, p. 18) use the idea of increased directionality as a way of discussing partially systematic/relational processing. For instance, such representations can be conceived of as multi-directional (rather than omni-directional) and the analogous suggestion is made that associative networks could be developed into (relational) tensor-product networks through changes in connectivity.

Cascade-correlation models

Cascade-correlation (Fahlman & Lebiere, 1990) has been proposed as a model of RR by Brook (1993) and Shultz (1994). (These proposals are covered in detail in Chapter 4 and so the discussion here is relatively brief.) Cascade-correlation builds a hierarchical, multilayer network structure over the course of learning, alternating between error-driven and correlation-driven learning. The latter takes place off-line after no further improvement can be made through error-driven learning. Cascade-correlation thus clearly corresponds to the basic structure of RR, and in being a constructive architecture facilitates incremental learning. The architecture has also been used in several process models of qualitative change during development reviewed in Shultz et al. (1995).

3.8.3 Connectionist–symbolic hybrids

Clark and Karmiloff-Smith (1993) divide the computational approaches they consider as potential RR models into 'connectionist–symbolic hybrids' and extended connectionist models which simply use 'more of the same' in the attempt to transcend the limitations of first-order connectionism.

The numerous references in the RR literature to qualitatively different representational formats (in turn supporting different degrees of flexibility and accessibility in somewhat the same way as a series of programming languages) might be taken as suggesting that the appropriate modelling framework for RR might be a connectionist–symbolic hybrid.

Clark and Karmiloff-Smith (1993) consider several examples of such models, noting that although such models do gain the resources of classical AI for representing structured knowledge, at the same time they lose the natural generalisation abilities of standard connectionist models, although this may relate particularly to those cases in which connectionism is used to implement (graded versions of) conventional AI architectures — compare the rule-extraction scheme discussed by Clark (1993a, p. 153), in which rules play a supporting role to a core of connectionist processing.

Clark (1993a, p. 152ff) also considers examples of hybrid models in the weaker sense, in particular commending an example of a distributed–localist connectionist hybrid on its representational multiplicity. He is however also critical of such approaches on the basis that even they involve much human intervention in the form of parameter-setting and, more importantly for potential developmental models, 'provide little indication of how such a [representational] multiplicity might automatically be developed by a system on the basis of a set of training inputs and connectionist learning rules. (p. 154).

Although not committing themselves either to hybrid or extended connectionist models, one of Clark and Karmiloff-Smith (1993)'s concluding observations is that:

> Finally we note that systems capable of providing these benefits will probably need

> to resemble the RR model even more closely than the hybrid systems described above in one further respect. They will probably need to generate multiple levels of increasingly abstract and manipulable representations of the basic knowledge the acquire *by connectionist means*. ... Current systems are prone to fall back on human intervention to bridge [the] gap [between statistically based, fully interwoven connectionist representations and single highly abstracted symbolic forms], [b]ut since the goal is a system that is self-driven to automatically generate the more symbolic representational forms, such interventions cannot be tolerated for long.
>
> (Clark & Karmiloff-Smith, 1993, p. 514, my emphasis)

while Karmiloff-Smith (1992b) observes that with time the boundaries of the class of hybrid models may well become blurred, and perhaps in the modular, mixed-learning-algorithm, and extended connectionist proposals for RR models, we may be seeing the beginnings of such a trend, although Clark and Karmiloff-Smith still seem to favour 'pure' extended connectionist approaches overall.

Greco and Cangelosi (1996b, p. 11) regard the issue of whether RR is best modelled using a (connectionist–symbolic) hybrid implementation as a pseudo-problem. They emphasise instead the question of whether symbolic representation is necessary, which they consider it is for introspective awareness, while 'composable elements, that act *like* words' are also necessary to a model of the RRH in their view.

3.8.4 Non-connectionist suggestions

Although, (presumably following the lead set in Clark and Karmiloff-Smith (1993), Karmiloff-Smith (1992b), Clark (1993a)), most proposed RR models have involved connectionism in some way, a few proposals from other computational modelling paradigms have also been made.

Kuscu (1993) suggests that a genetic-algorithm-based classifier system could provide a model of RR, although this proposal has not been explored. Classifier systems (see Forrest (1991) for instance) can be regarded as a kind of subsymbolic production system, incorporating condition–action rules (or *classifiers*), a credit-assignment strategy and a genetic algorithm. Such systems are proposed as models of the RR process on the basis of the kind of rule-transformation and consolidation which occurs over the course of learning/evolution, presumably uniting some of the advantages of self-modifying production systems, while avoiding the problems of rule-granularity and explicit representational framework noted above.

Grush (1994) suggests that RR effects may simply be a subset of those arising from the task of learning emulators and controllers, and suggests that the hypothesis could be fruitfully formulated in control-theoretic terms. Although his comments do not constitute an implementational suggestion as such, simulations of such constructs have been built.

Rutkowska (1994b) suggests that RR may be best understood as a computational process acting on action programs in a situated agent, and goes on to claim that attempts to provide a computational model of how viable activity patterns can become permanent adaptive changes need to concentrate on situated agents if they are to be of significance to an epigenetic approach.

3.9 Previous implementational work

Models of redescriptive effects have been constructed by Thornton (1995) and Greco and Cangelosi (1996b) in the domains of mobile-robot pursuit/evasion and a task involving the coordination of labelling and categorisation of stimuli respectively. These models both make use of competitive learning techniques to extract features deriving (partially or wholly) from previous training.

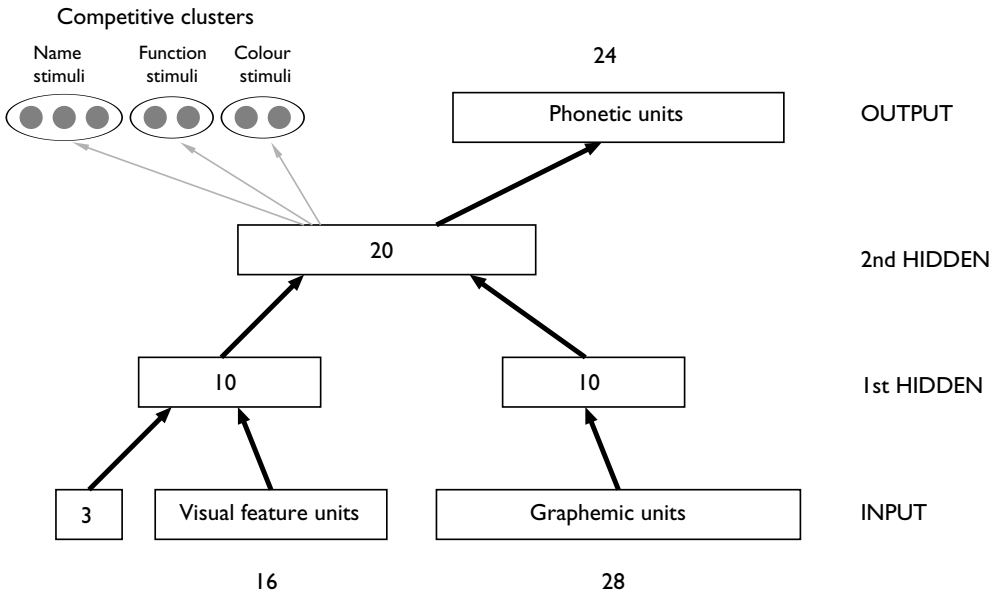


Figure 3.1: Greco and Cangelosi's redescription model. Connections in black are trained using error-driven learning (backpropagation), connections in grey using competitive learning after the other connections have been trained and frozen.

3.9.1 Greco and Cangelosi's model

Greco and Cangelosi (1996b) present a model of redescription in which competitive learning is used to map the results of training in an error-driven network onto a set of explicit categories. This simple explicitness of semantic information is used as the main criterion of redescription here, and Greco and Cangelosi define the act of making knowledge explicit as that of activating 'a local symbolic output corresponding to the category ... being named.' (p. 8).

Method

The error-driven learning was performed using a backpropagation network. This network was trained on a task which required it to coordinate the names and categorisations of objects, represented at the input as sets of localist features (see Greco and Cangelosi (1996a)). The input–hidden weights of the network were then frozen, while the hidden-output weights were replaced by connections from the hidden layer to a module of three banks of output units. These new hidden–output connections were then trained using competitive learning. The architecture of the model is shown in figure 3.1.

Results showed that this enabled the network to exploit the previously acquired semantic structure implicit in the hidden representation, giving rise to a structure similar to that shown by a cluster analysis of these representations.

Like Bechtel (1993), the authors enter into some debate about the significance of the new output units in a complete learner. The other implementation to be considered here shows that feature-detectors learned competitively can be re-appropriated into a complex and representationally rich system.

3.9.2 Thornton's Explicitation model

Thornton (1995)'s hybrid explicitation technique is not put forward as a model of RR as such (particularly in the sense of the RR model), although there are broad similarities (discussed below).

As mentioned in section 2.5.4, Thornton relates I-level representations to relational learning problems (i.e., problems such as parity whose solutions are not evident from the statistical profile

of the input data), while explicit representations reflect simpler mappings, which are explicit in the sense that they manifest themselves in the statistics. Explicitation works to transform implicit into explicit, in the sense that it brings non-statistical regularities within the grasp of (necessarily) statistical learners.

The network architecture is hierarchical and consists of multiple layers. These alternate between those with nodes trained using competitive learning and those consisting of variables extracted by a non-connectionist algorithm. Some modifications are also made to the conventional competitive learning scheme, in particular double-weighted connections are used which allow lower-order statistical effects to be exploited (Thornton, 1994).

Learning is incremental and phases of competitive learning are followed by a process of relational exploitation which forms a new layer of higher-level variables. These in turn support a new phase of competitive learning, and the whole process continues (ideally) until all statistical and relational effects have been fully exploited. The model is also conservative in that all inputs seen by the network are combinations of current and previous inputs.

The algorithmic explicitation technique works by identifying relationships in the data which involve constant distances. A set of data is considered to exhibit a relational effect just in the case that they can be 'arranged into a linear order such that each variable would show a constant difference from datum to datum' (Thornton, 1995, p. 10); something he also terms a *linear signature*.

Thornton interprets the internal feature-detectors formed through competitive learning in symbolic terms. The inputs to the network at any given time are either raw sensor data or 'symbols for more-or-less abstract and in most cases dynamic features of the current environment' (Thornton, 1995, p. 20). The network is thus seen as a dynamic multi-level recoding of environmental processes and events.

He compares this process with Piagetian concepts of change in a similar way to Shultz et al. (1995). The net is described as assimilating sensor data to its current internal representations. Structural accommodation involves the creation of new nodes, variables and layers, while non-structural accommodation corresponds to changes in network weights.

The explicitation procedure is also described as incrementally constructing 'a sequence of redescriptions or recodings of its sensory environment' (Thornton, 1995, p. 20) reminiscent of that in Karmiloff-Smith's conception of RR. However a limitation of the model as it stands, noted by Thornton, is that it relies heavily on scaffolding in form of a curriculum of learning scenarios provided by the (experimenter-controlled) environment.

3.10 Summary

Computational modelling provides cognitive scientists and developmental theorists with a dynamic testbed for hypotheses about qualitative change, as well as enforcing to a degree the discipline of a specification in computational terms, whether as a set of productions or the input and output representations of a connectionist network.

Although production-system models can provide fine-grained fits to experimental data, and have increasingly incorporated facilities for graded and incremental learning, the inherent explicitness of the underlying rules makes them seem inferior to connectionist systems in terms of capturing the progression from implicit to explicit in RR.

Connectionist schemes were defended as models of development on the basis that the underlying grain of quantitative change in such models was suited to sustaining revealing simulations of diachronic phenomena. The emergence of qualitative change from this gradual progression within a unitary framework was also put forward as a distinctive quality of connectionist simulations. In this context it was argued that hybrid models did not contribute significantly to our understanding of developmental phenomena. Incremental learning, via techniques such as architectural manipulation or training-set phasing, was identified as an effective technique in

simulating some kinds of representational change during development, and the related idea of *scaffolding* a representational trajectory was introduced.

The three characteristics of RR outlined in section 2.1: accessibility, explicitness, and sources of knowledge were discussed in a connectionist context. There is a sense in that, by demanding these capabilities, representational redescription can be viewed as a challenge to connectionism, requiring a developmental progression from associative to systematic and transferable knowledge. It was found that explicitness could be usefully recast for connectionism in terms of a continuum of system-relative levels of accessibility. The controversial related issue of systematicity could be usefully viewed as a product of scaffolded development, rather than a prerequisite in the cognitive architecture as Fodor and Pylyshyn (1988) insist. The role of domain-general constraints and (with certain limitations) domain-specific constraints was also considered something which might be explored within a connectionist model.

It was also argued that connectionist models were able to address issues in the RRH concerning the timing of mastery, its relationship to redescription and the role of continued on-line processing and residual error.

Implementational suggestions for the RRH were reviewed. These were found predominantly to involve connectionism, (presumably following the lead set by Karmiloff-Smith (1992b) and Clark and Karmiloff-Smith (1993)). Although the qualitative differences between formats (as well as the use of computer-metaphoric language) in the RR model might suggest the use of connectionist-symbolic hybrids, these were criticised on the grounds that they move away from the natural advantages of connectionism, such as direct generalisation, and that they require a great deal of hand-intervention, seeming to tell us little about how qualitatively different representational formats can emerge from a connectionist system.

Although most proposals for RR models involve augmented or weak hybrid (mixed-strategy or modular) systems, Plunkett (1993) argues that standard schemes such as backpropagation already embody systematic representations which are explicit in the restricted sense presented in this chapter. The proposal that associative and tensor-product networks could be related is intriguing but is not a process model of redescription as it stands.

Examples of implementations of redescriptive models are united in their use of competitive learning to extract features in conjunction with another process, either of error-driven learning (Greco & Cangelosi, 1996b) or a non-connectionist algorithm designed to exploit relational effects exhibited by the data set (Thornton, 1995). These models, as well cascade-correlation suggest that the explicit copying of whole networks put forward by Clark and Karmiloff-Smith (1993) is unnecessary. It is argued that hierarchical constructive models with alternating learning modes such as cascade-correlation and the explicitation technique constitute the most promising general class of broadly connectionist proposals for RR models. Cascade-correlation is selected for investigation here primarily since it is both an incremental and a wholly connectionist scheme and has been used in several simulations of qualitative developmental change, both micro- and macro-developmentally. The following chapter discusses these points in more detail.

Chapter 4

The cascade-correlation architecture and representational redescription

Introduction

In this chapter, the cascade-correlation architecture is presented more formally and in greater detail. The promise of the architecture as a model of RR is then discussed. Chapters 5 and 6 describe the experiments using cascade-correlation to model plurifunctionality and sequence-learning.

4.1 The cascade-correlation architecture

The cascade-correlation architecture (Fahlman & Lebiere, 1990) (Figures 4.1(a)–(e)) is a multi-layer supervised connectionist learning scheme. The most important difference between it and other multilayer schemes such as the standard backpropagation model, for instance, in that although it has (multiple) hidden layers, the number of hidden units is not predetermined. Instead these are 'recruited' (added) as necessary to the progressive reduction of error. The net begins with only the user-specified input and output layers (Figure 4.1(a)) and tries to learn the task using an error-driven learning method (in practice the quickprop algorithm (Fahlman, 1988) is used for speed). If this fails it then enters a correlation-driven recruitment phase in which units are trained off-line to the outputs. In this candidacy phase (Figure 4.1(b)), a pool of new hidden units is created, having only (randomly-weighted) incoming connections from the inputs and any previous hidden units. These are then trained via a gradient-ascent scheme in order to maximise the value of S in the following expression:

$$S = \sum_{o} \left| \sum_{p} (V_p - \overline{V})(E_{p,o} - \overline{E_o}) \right| \tag{4.1}$$

(where p ranges over patterns, o ranges over the output units at which the error is measured) i.e., the correlation[1] between the hidden-unit activations, V, and the sum of E_o, the error at the output units. The best of these candidates is then installed in a separate layer and connected to the output units. A pool of hidden units is used to help prevent the installation of useless units, e.g., those for which training got stuck in local maxima during the candidacy phase.

Since it is the absolute values of the correlations which are summed in Equation 4.1, candidates attend only to the magnitude of these correlations. As Fahlman and Lebiere (1990) observe,

[1] As Fahlman and Lebiere (1990) note, in practice the covariance, which is what S in fact denotes, worked better. Following Fahlman and Lebiere's usage, the term 'correlation' will be used in this context throughout to denote the covariance-based measure S.

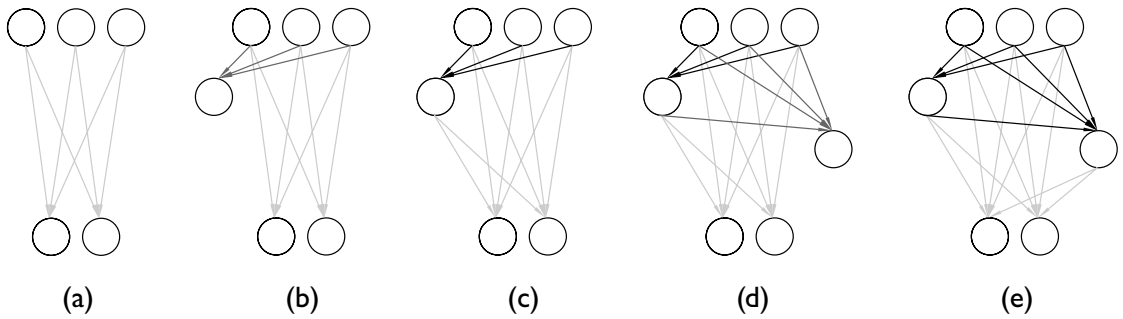


Figure 4.1: Learning phases in cascade-correlation over the course of two unit-recruitments. Arrows in light grey indicate connections which remain trainable using error-driven learning, those in darker grey indicate connections which are being trained using correlation-driven learning before installation, and those in black indicate connections whose weights have been frozen. Input units are shown at the top. For clarity only one candidate unit in each pool is shown.

if a candidate correlates positively with the error at a given unit, it will develop a negative weight in an attempt to compensate for that error, while if the correlation is negative, the weight will be positive.

On installation, the weights on the connections to the new hidden unit are frozen (Figures 4.1(c) and (e)) and the input–output and hidden–output connections are re-randomised and re-trained to readjust overall performance. This process is repeated until either the total error at the output has dropped to an acceptable level, or until any of various user-determined limits on epoch numbers have been reached.

Each learning phase, whether error-driven (Figures 4.1(a), (c) and (e)) or correlation-based (Figures 4.1(b) and (d)), continues until either the network has converged or there has been a lack of significant (proportional) improvement in error over a period of a number of epochs. The latter situation is known as *stagnation* and the parameter specifying the number of epochs is called the *patience*. Upper limits are also set on the number of epochs of each kind of learning which may take place in one phase. Stagnation causes the network to begin a new learning phase of the opposite kind to the current one, i.e., to pass from error- to correlation-driven learning or vice versa. Corresponding to the two kinds of learning is a division of sets of weights — *output-side* weights (shown in light grey) are trained using error-driven learning, while *input-side* weights (shown in dark grey/black) are trained using correlation-driven learning.

4.1.1 Performance characteristics of cascade-correlation

This section gives an overview of the performance characteristics and parameters of cascade-correlation particularly relevant to the modelling of developmental dynamics and conceptual change, summarising the main qualitative and quantitative differences between its performance and that of comparable multi-layer supervised learning techniques, in particular backpropagation (Rumelhart et al., 1986).

Learning speed

On a number of benchmark problems reported in Fahlman and Lebiere (1990), cascade-correlation performed significantly better than standard backpropagation. According to Fahlman and Lebiere these speedups are due to the following factors:

- Unlike backpropagation or quickpropagation (Fahlman, 1988), CC requires only a forward pass (rather than both a forward and a backward pass) through the network
- In cascade-correlation, many of the training epochs are run while the net is smaller than its final size; since the weight-values for the earlier, frozen, layers of the network do not

change, these can be cached, thus avoiding unnecessary calculations.

- Cascade-correlation uses freezing of existing structure and the restriction of each recruitment to a single unit. This is an attempt to combat what Fahlman and Lebiere (1990) call the *moving target* problem. Under these restrictions, the network only sees, a relatively fixed aspect of the problem, and is thus able to focus on it.

Incremental learning

As Fahlman and Lebiere (1990) note, cascade-correlation is well suited to incremental learning, i.e., in their terms, 'when information is added to an already-trained net.' (Fahlman & Lebiere, 1990, p. 11) (its suitability for capturing the related idea of incremental learning associated with developmental modelling is discussed in section 4.3.1). One reason for this is that the freezing of earlier-generated structure means that any feature detectors it embodies, once formed, are never cannibalized. Of course the extent to which these frozen sets of incoming connections are actively used by the network as feature-detectors depends on the strength of the weights formed between the hidden and output units. For instance a change in training set can cause these to change such that the effect of some of the previous input-side structure is diminished or lost. However the input-side weights have a strong mediating effect on the connections trained through error-driven learning, and Fahlman and Lebiere note that if the training set is changed, the output-side weights 'are quickly restored if we return to the original problem' (Fahlman & Lebiere, 1990, p. 11).

The constructive scheme used in cascade-correlation is also reminiscent of models and accounts inspired directly by biological development. For instance, Linsker's influential model of the development of the visual system made use of a scheme in which (self-organising) layers were added incrementally until the required higher-order feature-detectors had been formed. Quartz and Sejnowski (forthcoming) cite cascade-correlation as an example of a scheme which accords with their almost entirely constructive account of both neural and cognitive development.

Effects of parameters

Patience Patience controls how long the network takes to reach stagnation. i.e., for the proportional improvement in either an input- or output-side learning phase to fall below a certain level. Squires and Shavlik (1991) found this parameter to have one of the most important effects on the performance of cascade-correlation. By preventing overfitting, setting patience to a relatively low value can improve generalisation.

Input- and Output Epoch limits These parameters control directly the maximum number of epochs spent in any one phase of input- or output-side learning respectively. Like patience they can also be used to improve generalisation by preventing overfitting.

Size of candidate pool The size of the candidate pool controls the space searched for new feature detectors to freeze and install into the network. Although this paramter has not been used in previous developmental simulations, it is suggested here that it corresponds to a way of controlling the number of hypotheses which are generated about higher-order regularities in the data,

Generalisation

One of the most major problems besetting cascade-correlation is that of poor generalisation. According to Squires and Shavlik (1991) and Mohraz and Protzel (1996), both of whom have investigated the influence of freezing hidden-unit input weights on cascade-correlation (or in the case of Mohraz and Protzel on an architecture with similar connectivity pattern and training regime), it is this freezing which gives rise to the poor generalisation performance (and in some cases poor training-set performance) of CC. The reason given for this is that freezing can sometimes lock the network into regions of the solution space from which it cannot escape, either to converge on the solution, or in the case of generalisation, from an over-exact fit to the training set.

4.2 Comparison of other constructive algorithms with cascade-correlation

Cascade-correlation is an example of a class of connectionist algorithms known as *constructive* (or *generative* (Shultz et al., 1995)) algorithms. Other examples of such algorithms include Gallant (1993)'s pyramid and tower algorithms, tiling (Mézard & Nadal, 1989), and upstart (Frean, 1990). Gallant (1993) describes the motivation behind such algorithms as being 'to transform the hard task of building a network into the easier problem of single-cell learning' (p. 195). These algorithms are characterised by the following general features:

- progressive addition of structure (layers) over the course of learning
- addition of a single unit in each layer (CC, Tower, Pyramid)
- freezing of weights in previously added structure (Tower, Pyramid and CC)
- connections to all previous layers (CC and Pyramid only)
- option not to add a new node if doing so would be detrimental to performance (Tower and Pyramid only)

In terms of providing a model of RR, two of the features that cascade-correlation lacks might seem worth investigating. Firstly it is possible that a strictly layered network such as the one produced by Tiling might capture the progression through phases better (although its capacity to return to previous strategies might be less than that of a network which connects each layer to all previous layers). Secondly, the option not to add a node which would result in a performance decrement could be investigated as a control in the modelling of those domains in which RR exhibits a U-shaped behavioural curve (although this would bias the network towards error-driven feedback in terms of recruiting structure as well as in learning, which seems to move away from the idea of redescription as outside the main error-driven input–output mapping). Cascade-correlation is the only one of the above algorithms to have been used in model of human development to the author's knowledge (see 4.3.2 below for a review of these models).

The FlexNet scheme

Mohraz and Protzel (1996) present the FlexNet scheme. This is a family of models explicitly based on cascade-correlation, and extended (in a similar but more general way to that of Squires and Shavlik (1991)) by relaxing certain restrictions. Specifically, FlexNet differs from standard cascade-correlation in allowing variable numbers of units in hidden layers, variable degrees of cross-connection, variably sized candidate-pools and the freezing of weights to be made optional. Hidden units can also undergo candidate training in the context of existing hidden layers as well as new ones.

FlexNet also allows three connection strategies: *adjacent* in which only adjacent layers are connected as in a standard multi-layer perceptron, *full*, in which all units have direct connections to all others except those in the same layer, and *medium*, in which units are connected to the output units and all units in previous layers. Like cascade-correlation, FlexNet also uses two training phases, described as 'main' and 'candidate' (although the training scheme for the candidate units is not stated in Mohraz and Protzel (1996)).

In support of these extensions, Mohraz and Protzel claim that the restriction to single-unit layers in cascade-correlation is detrimental to generalisation. They also find that freezing slows training but does not affect generalisation, contrary to the findings of Squires and Shavlik (1991).

In terms of modelling redescription, the FlexNet model provides a superset of the features of cascade-correlation, and would thus allow manipulation of certain features such as freezing and unit-recruitment strategy.

To sum up, cascade-correlation represents a relatively powerful and general example of an important class of constructive, internal resource-phasing network architectures, and is the only

one which has been used to model development (although it has not previously been applied to the RRH). A more recent extension to cascade-correlation — FlexNet — provides a framework for varying certain aspects of the constructive scheme in cascade-correlation and would be an interesting tool for extending the work presented in chapters 5 and 6.

4.3 The promise of cascade-correlation as a model of RR

The relevance of using cascade-correlation in the construction of models of qualitative change during cognitive development is attested to by the results reported in Shultz and Schmidt (1991), Mareschal and Shultz (1993), Shultz (1991), Shultz, Buckingham, and Oshima-Takane (1994), Buckingham and Shultz (1994), Shultz et al. (1995). The relevance of these studies to modelling RR is discussed in more detail in section 4.3.2.

4.3.1 Cascade-correlation as a model of RR

Cascade-correlation has also been proposed (Brook (1993, 1995), Shultz (1994)) as a possible connectionist model of representational redescription. In support of this, both authors cite the fact that it is an example of an incremental learning scheme, which makes it a suitable candidate for developmental modelling in general, and that in unit-recruitment it incorporates a natural mechanism for supporting qualitative change. More specifically to cascade-correlation, the changes in performance it passes through are both conservative and hierarchical, and involve alternations between phases of differently focused types of learning. The transitions between these phases are also prompted by the net's achievement of a stable state. These aspects are considered in turn in the sections that follow.

Karmiloff-Smith (1994, p. 739) has responded that cascade-correlation 'seems to capture the first level of redescription in ways that are very close to the intuitions underlying RR' (although she does not elaborate on this perceived correspondence), and also comments that she finds illuminating the 'notion that cascaded hidden units afford the construction of increasingly powerful knowledge representations that were not available to developmentally earlier instantiations of the network' (p. 738). She also now agrees that it is possible for such models to provide a framework for studying developmental principles at a level more general than that of individual tasks.

Conservative changes in qualitative performance

Cascade-correlation preserves both previous structure and (partial) solutions in the frozen incoming weights to the hidden units. This gives the network the potential to return to previous solution states at least over the immediately succeeding unit recruitments. Shultz and Schmidt (1991) report that in their model of the balance-scale (torque-difference) task, around the time of a qualitative change in behaviour the network tended to go through a period where it would alternate between strategies.

The cascade architecture also resembles the hierarchical nature of the representations proposed by the RRH and means that any new learning which takes place must happen with respect to all earlier, frozen, hidden structure.

Incorporation of alternating learning modes

Central to the RR model is the idea that learning within phases has a different focus to the process which moves the learner between phases. Specifically, learning within phases is driven by the need to make quantitative performance improvements, while the (redescriptive) process which brings about phase transitions is not directly error-driven and acts to bring about greater flexibility and explicitness. Cascade-correlation also incorporates two alternating learning modes, and like RR, one is error-driven and acts directly to try and improve performance while the other is indirect, occurs off-line, and is associated with biasing a (microscopic) quantitative change in processing power in the form of the recruitment of a new one-unit hidden layer. Each learning-mode change is also prompted by the net's having reached a stable state in the previous mode,

and this can be seen as corresponding to the stable states of (at least partial) task-mastery required for redescription in the RRH.

To the extent that the motivations for redescription are open to question, cascade-correlation can provide a framework within which to investigate such issues.

Incremental learning and developmental modelling

Incremental learning (in the context of developmental modelling) is a broad term which refers in connectionist theory to the idea that learning a task is best (or only) achieved through the 'staging' of learning, usually through manipulation of resources (or *resource phasing*). These resources may either be internal to the network, such as number of internal nodes or length of attentional window in recurrent nets (Elman, 1990b, 1991), or external, such as ordering or constituency of training set. Resource-phasing works to facilitate the learning of complex tasks by making use of the fact that the initial inability of the network to deal with complexities, such as long-range dependencies for instance, can allow it to focus on the simpler aspects of the same problem. Thus resource-phasing may act to produce a conservative problem decomposition (Clark & Thornton, 1993), the simpler aspects of the problem acting as building blocks on which subsequent, more complete, solutions may depend (Clark, 1993b) (see 3.5 for further discussion).

Incremental learning also appeals to modellers more interested in capturing the path or trajectory through the process of learning of a task than in eventual successful performance in itself. For instance, resource manipulation allows the trajectory through the learning process to be actively shaped using general means without the need to resort to hand-crafted or synchronic modelling, and thus to move away from the emergent dynamics of a network model. The hidden-unit-recruitment strategy of cascade-correlation is an example of resource-phasing, each step serving to increase the power of an initially very limited network.

Fahlman and Lebiere (1990) refer to the process of hidden-unit recruitment as adding 'higher-order feature detectors to the network'. The idea of a hierarchy of feature detectors is central to the general approach to problem decomposition presented in Clark and Thornton (1993). The question of whether, and under what conditions these two ideas of feature detectors are to be distinguished from each other, remains to be explored.

Clark and Thornton (1993) also claim that:

> More generally ... any cascade of processors in which the upstream devices sort, filter or recode incoming data holds out the promise of promoting successful learning
>
> (Clark & Thornton, 1993, p. 40)

by guiding the network towards the kind of early learning which reduces the statistical complexity of the task. Clark and Thornton give the following list of recoding strategies used in recent incremental learning models: evolved pre-processors, acquired, training management, initial short-term memory limitation, and modular decomposition techniques. Cascade-correlation inherently implements initial short-term memory limitations in in that its architecture starts off with minimal power — in the recurrent version for instance, the initial lack of hidden units means that the net begins with no recurrent elements at all. In its building of a hierarchy of feature detectors, CC can be also be said to implement acquired recodings. Training management is not inherent to CC, but is compatible with it and has been used by Fahlman (1991) in the Morse-code learning example with RCC. It would also be possible to install hard-wired unit structure to play the role of innate or evolved pre-processors. Recurrence may also be seen as a domain-general constraint on processing (Karmiloff-Smith, 1992c).

Although Clark and Thornton do not commit themselves to whether the cascade of recodings or filters in cascade-correlation are likely to perform the right kind of reductions in complexity to allow relational tasks to be learnt, they note both that single networks rely on decompositions of a single task to constitute a trajectory (p. 38) (i.e., tend to be confined to learning

tasks which have a conservative incremental solution). They also consider (p. 22) that even if sophisticated learning algorithms such as cascade-correlation are able to solve 'hard' problems such as parity, these inevitably introduce restrictive assumptions about the nature of hard problems in terms of which recodings are possible.

However, unlike single-network schemes such as backpropagation or SRN's, central to cascade-correlation is the preservation of the results of previous learning in the form of the frozen input–hidden and hidden–hidden weight structure. Although use of this mechanism does not avoid the problem of interference between tasks learned in sequence by a single network, it does suggest at least the possibility of building early feature detectors via training on other tasks, rather than necessarily restricting training to subsets of a single task. The experiments on transfer of simple counting-related knowledge in chapter 6 provide a partial investigation of the use of scaffolding on structurally related tasks in conjunction with cascade-correlation.

In incorporating a standard error-driven learning algorithm, cascade-correlation inherently shares some of the general qualities which are considered by Karmiloff-Smith to make connectionist systems appropriate models for redescription, in particular at the initial I level. These include gradualistic, distributed representations, and adjunction of representations (see 3.2.3 for discussion).

However, Karmiloff-Smith (1994) expresses several reservations concerning cascade-correlation. One of these is that she sees the fact that the move from stagnated error-driven learning to correlation-driven learning is automatic as implying that behavioural mastery is sufficient for redescription in addition to being necessary, as Karmiloff-Smith often claims. Karmiloff-Smith (1994) thus seems to view both kinds of learning in cascade-correlation as taking place within the input–output mapping, preferring to consider redescription to be something which acts independently of this.

4.3.2 Previous developmental models using cascade-correlation

As noted above, a number of previous developmental models have been constructed by Shultz and his collaborators, although none of these has dealt explicitly with capturing the RR model. Shultz et al. (1995) provides a survey of this work, while the following sections focus on two models particularly relevant to the work presented in chapters 5 and 6.

Seriation

As mentioned in chapter 3, Mareschal and Shultz (1993) used cascade-correlation to construct a model of development on the seriation task (Inhelder & Piaget, 1969). Their model involved a modular arrangement consisting of two networks, dealing separately with the tasks of which stick to move next and where to move it. The network was trained to perform according to the operational method (i.e., select the smallest, as yet unordered element and move it into its correct position in the series).

This model was able to capture the progression through a series of strategies associated with this task. Analysis of the representations formed during learning using Hinton diagrams revealed further that the network had achieved the correct, staged, solution in a gradualistic manner which contrasted with the solutions formed in previous cascade-correlation models such as that of Shultz and Schmidt (1991). This leads Mareschal and Shultz to comment that in general '[s]ome representational changes appear to require qualitative changes in representational power, whereas others do not.' (p. 5).

Seriation is also cited by Karmiloff-Smith (1992b) as a canonical example of a sequence-learning domain in the sense addressed by the RRH, (although a detailed account within the RR framework is not presented). This model exhibits the pattern of ends-inwards success associated with sequence learning by the RRH. As Mareschal and Shultz (1993) note,

> The Hinton diagrams also revealed that the building of a representational structure in the network began by adjusting weights leading to those dealing with the short

> end of the series and was progressively extended along the length of the series until finally appropriate weights were found for those units coding the larger end of the sequence. This is consistent with [the] suggestion that children build a linear order mental representation of the seriation task by proceeding from the ends of the series inwards.
>
> (Mareschal & Shultz, 1993, p. 5)

Children's acquisition of first- and second-person pronouns in English

Shultz et al. (1994) used cascade-correlation to model the acquisition of first- and second-person pronouns in English. The phenomena to be modelled included the so-called reversal errors temporarily made by some children. These errors involve treating the pronouns 'you' and 'me' as if they were names and thus always had the same referent, with the consequence that others are addressed as 'me', and self as 'you'.

Their model made use of training-set phasing in conjunction with cascade-correlation in order to investigate the roles of directly addressed and overheard speech. The study also involved an investigation of the effects of changes to the composition of the training set on further learning (something they termed. 'second-phase' training). These changes involved a change in token-frequencies and the proportion of child-addressed speech.

This study has direct relevance to the RRH in that the phenomenon modelled is cited by (Karmiloff-Smith, 1992b) as an example of the effects of explicitised representations on behaviour as well as having some simlarities to the study presented in chapter 5, particularly the simpler form discussed in section 5.4. Shultz et al.'s model is the only previous study of cascade-correlation in simulating a psycholinguistic phenomenon known to the author.

Shultz and Schmidt (1991) also used cascade-correlation to model the balance-scale (McClelland, (1989), see also section 3.3.1), capturing all four of the strategies involved, including the minor regressions observed in the task.

Shultz et al.'s models and Piaget

Shultz et al. (1995) also compares the alternation of the two types of learning process in cascade-correlation to Piaget's ideas of assimilation and accommodation, arguing that it is thus possible to view cascade-correlation as a model of equilibration. Despite this direct relation of the architecture to Piagetian terminology, there is nothing inherent in it which implies that the qualitative shifts it exhibits should be seen as age-related stages for instance, and it is not thereby compromised as a potential model of RR.

4.4 Summary

This chapter has presented the cascade-correlation architecture in more detail, surveying its practical strengths and weaknesses and comparing it briefly with similar constructive network architectures. It was argued that cascade-correlation is promising as a model of representational redescription as it provides conservative qualitative change, alternating learning modes, and supports incremental learning. It also incorporates the qualities of standard (error-driven) connectionist learning such as gradual change, graded and distributed representations and the representational adjunction characteristic of level I of the RR model. Previous cascade-correlation models of qualitative developmental change in the domains of child language acquisition and the Piagetian balance-scale and seriation tasks were surveyed and presented as evidence that cascade-correlation is well-suited to capturing qualitative developmental change.

Chapter 5

Cascade-correlation as a model of RR in plurifunctionality tasks

5.1 Introduction

This chapter describes experiments carried out using cascade-correlation to model redescriptive effects in the comprehension of the French article system (Karmiloff-Smith, 1979a, 1992b). In this setting redescription involves the formation of a linguistic subsystem relating the common features of the articles, from an initial state in which representations are stored individually. This micro-domain was chosen as it provides a good example of a situation conforming to the three-phase RR model.

5.1.1 The playroom experiments — comprehension of the article system

The following simulation is based on a series of experiments originally designed to test children's understanding of the article system in French (Karmiloff-Smith, 1979a, p. 171 ff.) but also put forward by Karmiloff-Smith (1986, 1992b) as an example of how knowledge of articles is redescribed into a linguistic subsystem of which the children show increasing awareness.

The experiments involved an array consisting of two playrooms, in which a boy and girl doll were placed — see figure 5.1.

In each playroom a selection of one or more of several kinds of common object was placed. The child taking part in the experiment is told that the dolls will be given groups of toys to play with (which will be changed after a few games). The task is to guess whether the experimenter is speaking to the boy or the girl doll each time they ask 'Lend me a/the X', based on the contents of each doll's playroom. The expected response is to choose the playroom containing a single X when 'the' is used and more than one X when the article is 'a'. For example in figure 5.1 the question 'Lend me a W' would most naturally be addressed to the boy doll, while 'Lend me the W' would be addressed to the girl doll, because there is only one W in the right-hand playroom and several in the left-hand playroom.

Karmiloff-Smith (1979a) found that children of all ages successfully chose the room containing a single instance of a particular object when the definite article was used. For the indefinite article however, while the youngest children studied (3-year olds) gave the appropriate response in 90 percent of cases, this had dropped by age 5 to as little as 29 percent, and it was only by the age of 8 that performance started to approach its original high level. Figure 5.2 shows the results of the original experiment.

Associated with these three phases are symptomatic qualitative changes in the way in which the children justify their choice of addressee: younger children justify their answers (often erroneously) with reference to real-life situations, middle-age children refer to the situation in the array, while older children are able to discuss the linguistic conventions at work and emphasise

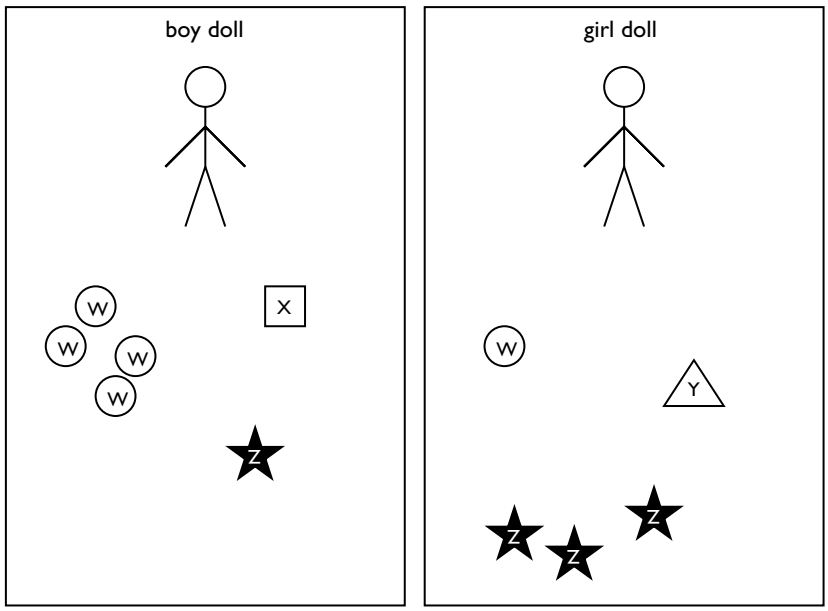


Figure 5.1: The experimental array used in the playroom experiment (after Karmiloff-Smith (1979a, p. 65))

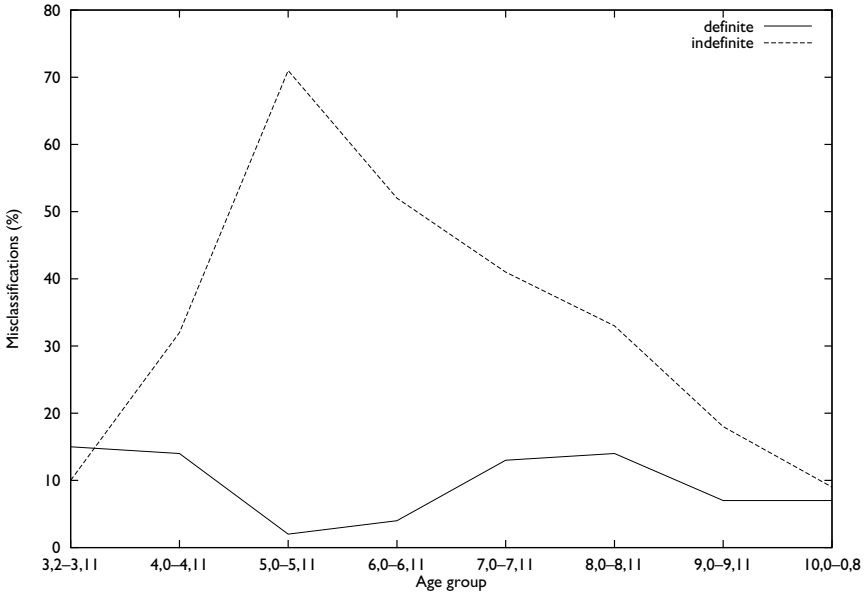


Figure 5.2: Percentage misclassifications on the playroom comprehension task. The drop in performance on the indefinite article (corresponding to the rise to 70% error here) in the middle age group is striking, while performance on the definite article remains relatively good throughout. (From data on Karmiloff-Smith (1979a, p. 175))

the article used by the experimenter (because, in the terms of the RR model, they now have explicit conscious access to the linguistic subsystem underlying their performance).

The decline in performance in the middle group is thought by Karmiloff-Smith to be due to the fact that in French the word-form indefinite article (un/une) also functions as the numeral one. While the indefinite article usually has a non-specific function (corresponding to the usage of 'a' or 'any' in English), its numeral function is more specific and there is thus a potential conflict between the two functions. The subjects' interpretation of the definite article appears to remain correct throughout, although, as the RRH would predict, it has a different status according to the age of the children.

The changes in interpretation of the indefinite article are accounted for as follows by Karmiloff-Smith:

> With regard to the indefinite article, it is suggested that the youngest children interpret it correctly because of the absence of a focus of attention (which the definite article implies for them). The interpretation of the 5 to 7 year olds seems to stem from concentrating on a new function conferred on the indefinite article: that of being a numeral. ... From 7 years, conflicts between the numeral function and the non-specific function of the indefinite article became apparent. [through the addition of markers in production] Over 8 years olds interpreted the indefinite article as a non-specific reference [and epilinguistic data they produced suggested] an understanding of non-specific reference.
>
> (Karmiloff-Smith, 1979a, p. 184)

So the basic behavioural pattern to be modelled is a U-shaped curve in performance on comprehension of the indefinite article (particularly in its non-specific function). In terms of redescriptive effects we would expect this to be accompanied by, and symptomatic of, an increasing systematicity in the underlying representations.

5.1.2 The changing status of articles — from unifunctionality to plurifunctionality

The U-shaped behavioural curve observed in the playroom experiments is explained by Karmiloff-Smith (1979b) by the idea that the underlying representations of the article–function associations undergo qualitative change — specifically they become more systematic. Although the youngest children exhibit apparently adult usage and comprehension, associating the same article form with different functions, this is in fact comprised of the use of a collection of distinct article-form–function pairs (termed by Karmiloff-Smith 'unifunctional homonyms') produced by separate procedures conforming to a one-form–one-function constraint.

Karmiloff-Smith (1979b) claims that this change in the status of meaningful units from unifunctional to *plurifunctional* is an example of the kinds of spontaneous qualitative changes which occur as part of representational redescription.

The corresponding overmarking of the determiner and partitive functions of a word like 'un' (used in its non-specific sense) observed in the productions of middle-age children points to a new awareness of the dual functions of such items and the (temporary) perceived need to mark functions explicitly using separate words. The eventual return to adult comprehension and production reflects the fact that it is now possible for one form to serve more than one function, i.e., the form now has plurifunctional status for the child.

5.1.3 Plurifunctionality, the playroom experiment and RR

In terms of the RR model the explanation of this macro-developmental progression runs as follows (Karmiloff-Smith, 1992b, p. 57). The initial behavioural mastery exhibited by the youngest children calls on two independently stored level-I procedures which map phonological forms onto specific functional contexts. Although able correctly to produce and interpret articles in

this deictic context, what they do not 'know', claims Karmiloff-Smith, is that there is a functional relationship between these efficiently functioning procedures; nowhere is there an explicit (E1) representation of the functional links between the articles. After this initial behavioural mastery has been achieved, representational redescription acts on the separate procedures producing the unifunctional homonyms, (which remain intact and able to be called upon for certain purposes) to give explicit E1-level representations. This makes it possible to link the common phonological form across the two form–function pairs. This newly formed representational link explains the appearance of errors in the 5-year-olds' production and comprehension. In production these appear in the form of the overmarking of separate functions and in comprehension in the form of mistakes concerning which of the two functions of an article is intended. Finally, the differing ways in which children justify their responses verbally also supports an RR interpretation — only the oldest children could account for their correct performance by making reference directly to the linguistic subsystem at work. This would suggest that they were representing the micro-domain in E3 format.

The experiments reported here are concerned with modelling the comprehension errors and performance which appear at level E1, as well as capturing the overall U-shaped curve in performance.

5.2 Modelling the playroom experiment using cascade-correlation

5.2.1 Input representation

Input data are of the following form. The inputs are divided into three banks representing the array (the two playrooms), the object asked for, the article associated with that object in the question, and the function of the article.

The array is represented such that there can be either 0, 1 or 'more than one' (indicated by M) of each object for each addressee. Each object-type is thus associated with a pair of values, the first corresponding to the number of objects in the left-hand playroom (the boy doll's) and the second to that in the right-hand (the girl doll's). In the question-object bank a 1 indicates that that object is the one requested using a one-of-n[1] encoding. A 1 on the article unit indicates the indefinite article, –1 the definite article. The function bank again uses a one-of-n encoding over three units to give orthogonal encodings for the non-specific sense of the indefinite article ('a'), the definite article and the specific sense of the indefinite article ('one').
Using this encoding scheme the vector

$$\underbrace{\underbrace{\underbrace{(\mathrm{M},1)}_{W}\ \underbrace{(1,0)}_{X}\ \underbrace{(0,1)}_{Y}\ \underbrace{(1,\mathrm{M})}_{Z}}_{\text{array}}\ \underbrace{1\ \ 0\ \ 0\ \ 0}_{\text{question objects}}\ \underbrace{1}_{\text{article}}\ \underbrace{\textit{specific}}_{\text{function}}}$$

would represent the situation shown in figure 5.1, with the experimenter asking about the question 'Lend me a W'. The appropriate response is to say that the experimenter is addressing the boy doll, since it is the left-hand playroom which contains more than one W.

The encoding for the array was adapted from the experimental setup used by Karmiloff-Smith (1979a, pp. 171–2). In her experiment, children were asked a total of 16 questions relating to four contexts (p. 171) comprising differing arrangements of objects within the two playrooms. In each context four different kinds of objects were used, two of which appeared in differing quantities in each playroom, while the others appeared in one playroom only. In terms of the encoding used above, each arrangement of objects would be represented by a permutation of the multiset

$$\{(\mathrm{M},1),(1,\mathrm{M}),(1,0),(0,1),\underbrace{(0,0)\ \ldots\ (0,0)}_{11}\}$$

[1] a scheme in which each possible distinct data-item is represented by activating a single unit of n possible units (hence one-of-n), giving a set of orthogonal vectors

corresponding to the fact that there are 15 possible objects, 4 of which are used in any given context during the experiment.

A pilot study using an array containing all fifteen different input object-types had shown that there was a large overhead due to the network's having to learn the associations between the banks of inputs conveying information about objects, and that this became prohibitive as the number of distinct objects increased. Since the intention in the original experiment was that the objects chosen should already be familiar to the children, it was decided that this overhead of object recognition could be reduced without affecting the essence of the experiment. The fifteen different objects used in Karmiloff-Smith's experiment were thus collapsed into four, with each object appearing in a different 'role' (i.e., appearing in both playrooms or only one, or appearing more than once in a particular playroom).

The original experiment also controlled for differences (or lack of them) between individual objects within a group (e.g., colour), and these differences were also elided for simplicity of encoding.

Although the encoding is presented in its above form here for clarity, in practice, for each of the object-type pairs in the array bank and also the article function, a one-of-n vector was generated. This transformed representation was intended to avoid making any information in the inputs unnecessarily explicit. A more compact representation, in which the values shown as 0, 1 or M above were mapped directly onto the values –0.5, 0 and 0.5 respectively, was used in some pilot studies, but was rejected in that it rendered certain important relationships in the data explicit before training had even begun. The 0 targets also made Hinton diagrams difficult to interpret. [2]

5.2.2 Composition of training data

The network was trained to associate situations in which an object is asked for using the definite article with the playroom which contained exactly one instance of that kind of object. In the case of the indefinite article, the network had to learn to associate the single form with two different functions. The specific case required the same response as that for the definite article, i.e., the singleton object, while the non-specific case requires that the addressee selected should have more than one of the kind of object being requested.

Object configurations in training data were produced by initially selecting four pairs randomly with replacement from the set {(1,M),(M,1),(1,0),(0,1),(0,0),(M,0),(0,M)}, subject to the constraint that at least one of the pairs was not (0,0), i.e., that there was at least one object in the whole array. An article was then chosen randomly, and, if indefinite, a random choice of the specific ('one') or non-specific ('a') function was also made, while definite articles always took the same function value. Once the article/function had been chosen the array was then checked against it to see whether it contained at least one of the object-type pairs appropriate to the article–function pair. For instance, the indefinite article and non-specific function pair applies to situations in which one playroom contains more than one of a given object, i.e., is represented by one of the pairs (M,1) or (M,0) (or the reversed versions). If the randomly generated array contains more than one applicable object-type pair then one of these is chosen at random.

Certain situations are excluded, in particular the ambiguous situations in which both addressees have one or more than one of a certain object (this restriction was also made in the experimental setup in Karmiloff-Smith (1979a)). With these situations excluded, it is possible to determine the expected addressee in each case, and this is used as the target data at the output — a 1 on the output unit means that a sentence of the form represented by the inputs would usually be addressed to the girl and vice versa.

[2] Although clearly if two variables are used to represent data which can be represented using only one then their values will be anti-correlated, the resulting representation (of the whole input vector) be less inherently systematic than that formed using a single variable.

Article	Ambiguity	Array	Question objects	Response
definite	unambiguous	((0,1),(M,1),(1,M),(1,0))	(0,0,0,1)	room with 1 (left)
definite	ambiguous	((1,M),(0,1),(M,1),(1,0))	(0,0,1,0)	room with 1 (right)
indefinite	ambiguous	((M,1),(1,0),(1,M),(0,1))	(1,0,0,0)	room with M (left)

Figure 5.3: The three test situations in Karmiloff-Smith (1979a)

Training in this way with the intended function made explicit in every case was intended as pretraining corresponding in a broad sense to the previous linguistic experience of children in this microdomain. Karmiloff-Smith (1979b) notes that in daily discourse 'such ambiguity rarely exists due to contextual clues' (p. 95) and the explicit functions were intended to indicate such context.

Using a method similar to that used by Plunkett and Marchman (1993), weight matrices were saved after each phase of output-side learning,[3] giving one matrix for each hidden-unit configuration of the network, and tested on a data set not used in training to investigate the progressive systematicity of the representations formed within the network as well as generalisation.

5.2.3 Test data

In order to test the generalisation of the learning of the different semantics for the indefinite article over the course of learning, the experimental arrays and questions were again adapted from Karmiloff-Smith (1979a, pp. 171–2). In the experiment, children were asked a total of 16 questions relating to four contexts (p. 171) comprising differing arrangements of objects within the two playrooms. In each context four different objects were used, two of which appeared in differing quantities in each playroom, while the others appeared in one playroom only.

The experimental items in the original experiment were spoken with normal intonation, which implies the non-specific sense for the indefinite article. The test set thus also used the non-specific function for the indefinite article.

Again following Karmiloff-Smith (1979a), the three situations represented in the test data were: definite article with a type of object which was present in only one playroom (unambiguous), definite article with a type of object present in different quantities in each playroom (ambiguous), and indefinite article with an ambiguous array as in the definite-article case. Figure 5.3 summarises these categories, which will be referred to throughout. The role of the unambiguous situation in the simulations was slightly different to that in the original experiments — there definite unambiguous cases were presented first to test children's understanding of the experimental setup, whereas here they serve as a test of the pretraining on both object recognition and article–function mapping.[4] All 144 possible such exemplars were generated and used as the test set with the additional condition that none of these should appear in the training set.

[3] In fact, networks were restricted to recruiting one unit per trial and the defeated networks reloaded until the network reached criterial performance (or 'victory' in the context of cascade-correlation). Since the output-side weights are randomised after each recruitment, the network sometimes continued output-side training for some epochs after reloading. The numbers of output-side epochs immediately after recruitment tended to be small and both the overall epoch-numbers as well as the number of hidden-units recruited were comparable with a control network which was trained continuously (i.e., without saving and reloading of weight matrices).

[4] In hindsight, it would have been desirable to add unambiguous indefinite-article cases as an indication of baseline performance on the difficult indefinite non-specific category.

5.2.4 Overview of method

Network configuration

The networks used in these experiments were non-recurrent cascade-correlation networks with 32 binary inputs and one binary output unit[5]. In all of the experiments presented in this chapter exemplars were trained in batch, i.e., weights were only updated after the entire training set had been presented to the network. This is the usual mode of training for cascade-correlation many of whose control parameters are epoch-based. In each case the size and constituency of the training set were kept constant throughout training. Training continued until the training-set error reached 0 bits.

The parameter settings were either arrived at by hand during preliminary investigations or based on based on defaults given by Fahlman (personal communication, 1995). Experimental manipulations involving variations to certain parameters are discussed below.

Causes of qualitative change

Internal causes of change Cascade-correlation includes many parameters and this study focused on a small number of these which were considered particularly relevant to the production of qualitative change, in particular those shifts in representational style associated with RR.

This study focuses on two of these parameters: patience and candidate-pool size. Patience controls the number of epochs the network waits before giving up while improvement is proportionally small. This controls the timing of the point termed stagnation, at which a given phase of input- or output-side learning ends.

Varying the size of the pool of candidate units corresponds to increasing the space searched for possible feature-detectors based on the features of both the raw input as well as the recodings of hiddens downstream. A large pool size allows large changes in the magnitude of the correlation between the incoming weights to a newly recruited hidden and the residual error at the output unit(s). These weights in turn have a strong mediating influence on the outputs and thus the overall behaviour. During the experiments reported here it was found that weights between hidden and output units tended to be relatively high.

External causes of change Biases in the frequencies of certain types of exemplar in training data are acknowledged as contributing to the modelling of stagelike qualitative transitions in networks trained using supervised learning schemes (Plunkett & Marchman, 1993; Shultz et al., 1995; McClelland, 1995). Variations in the proportions of different classes of article–function pairs were thus investigated.

Analysis of behaviour

The main index of behaviour during training was the proportion of misclassifications on different categories of inputs (e.g., definite article, unambiguous situation) on both training and test sets as in Karmiloff-Smith (1979a). It was thus also assumed that a classification error implied a choice of the opposite case, e.g., for the indefinite article in the test set a misclassification implied that the specific (or numerical) rather than the non-specific function was intended — as Karmiloff-Smith (1979a, p. 176) notes 'it is the fact that the definite article is *not* used that is a clue to the more appropriate response'.

Analysis of representations

Hinton diagrams were used as the main means of examining internal representations directly. The limitations of this method are noted by Shultz and Elman (1994). In particular, such analyses can be difficult to compare even between different runs of the same network, and also do not take into account the sign or magnitude of the corresponding input signal.

As Shultz and Elman (1994, p. 1118) note, because the cross-connections in architectures such as cascade-correlation carry so much of the workload, applying statistical techniques such as principal components analysis (PCA) (often applied to the hidden layer of backpropagation

[5] except in one variant in which a simplified input encoding was used — see section 5.4

or simple-recurrent networks) to hidden-unit activations provides, at best, only a partial picture of the solution formed in the network. The main alternative method of analysis which has been proposed for cascade-correlation is contribution analysis (Shultz & Elman, 1994; Shultz & Oshima-Takane, 1994). This can provide an analogue to PCA for cross-connected networks such as cascade-correlation. However as Shultz and Elman (1994) note, it is unsuitable for use with binary input values such as those used here, which were chosen since the use of multi-valued input units was considered to be too representationally biased, as well as making weight-values difficult to interpret directly.

Training-set biases

The frequency with which children hear utterances using each form–function pair was not known. [6] Datasets containing differing proportions of exemplars were thus generated according to several kinds of scheme. Tables 5.1(a), 5.1(b), and 5.1(c) give the different configurations according to which the training sets were generated. Configuration A simply balanced the proportions of indefinite and definite article exemplars, balancing proportions of each sense and then situation (or ambiguity) within these. Configuration B had equal proportions of definite, indefinite non-specific and indefinite specific exemplars, again with situations equally represented within these. Configuration C was mainly intended to provide a bias towards the definite article, in the interests of investigating whether this would address the surprisingly poor performance on this category which had been observed in pilot studies.

5.3 Results

The main set of experiments used the input representation given in section 5.2.1. In order for the network to learn the correspondences between the two banks of units representing object-type information in the playroom arrays. Pilot studies had shown that several thousand exemplars were needed and the training sets in this section each consisted of 2000 unique exemplars.

5.3.1 Basic performance

The basic performance of cascade-correlation on the three dataset configurations is summarised in table 5.2. These results show that using input data restricted to four object-types the network was able to learn the basic task including that of matching object identities between the array and question-object banks.

5.3.2 Misclassifications

As noted above, misclassifications on different categories provide the main (behavioural) means of diagnosing qualitative change. The proportions of misclassified exemplars from the training and test sets were recorded each time a hidden unit was recruited.

Misclassifications on training set Figure 5.4 shows the misclassifications across the different categories of exemplars over the course of training for networks of each of the three configurations.

Fluctuations in the relative numbers of misclassifications in each category correspond to shifts in attention to the specific or non-specific function, or to groupings of objects, which are evident from Hinton plots of the same networks. For instance, the abrupt levelling off of proportional error observed between the recruitment of the second and fourth hidden units in

[6] This is presumably a result of the difficulty of collecting the appropriate data, which would need to record not only the frequency of article use in child-directed and child-overheard speech for a language in which the indefinite article had both non-specific/numeral function, but also the intonation or other linguistic markers which serve to indicate function. Although longitudinal studies exist which record parents' and caregivers' speech for French-speaking children (e.g., as part of the CHILDES project (MacWhinney, 1991)), to the author's knowledge no analysis of article-function frequency has been performed on such data.

Definite article	50%		ambiguous	25%
			unambiguous	25%
Indefinite article	50%	non-specific	ambiguous	25%
		specific	unambiguous	12.5%
			ambiguous	12.5%

(a) Training set balanced between definite and indefinite articles (Configuration A)

Definite article	33%		ambiguous	16.5%
			unambiguous	16.5%
Indefinite article	67%	non-specific	ambiguous	33.5%
		specific	unambiguous	16.75%
			ambiguous	16.75%

(b) Training set balanced between definite, indefinite article (non-specific sense) and indefinite article (specific sense) (Configuration B)

Definite article	67%		ambiguous	33.5%
			unambiguous	33.5%
Indefinite article	33%	non-specific	ambiguous	16.5%
		specific	unambiguous	8.25%
			ambiguous	8.25%

(c) Training set biased towards definite article (Configuration C)

Table 5.1: Configurations of proportions of different categories used in training data

Input Epochs	Output Epochs	Average Hiddens	Average Epochs	Min/Max
150	150	7.5 (6/8)	999	705/1276
100	100	7.3 (6/8)	898	660/1028
100	50	7.2 (5/9)	732	523/911
50	100	7.5 (6/9)	896	709/1074
50	50	7.5 (6/9)	710	570/826
20	20	9.8 (9/9.8)	415	372/430 †
10	20	defeated	–	–
10	10	defeated	–	–

(a) Configuration A

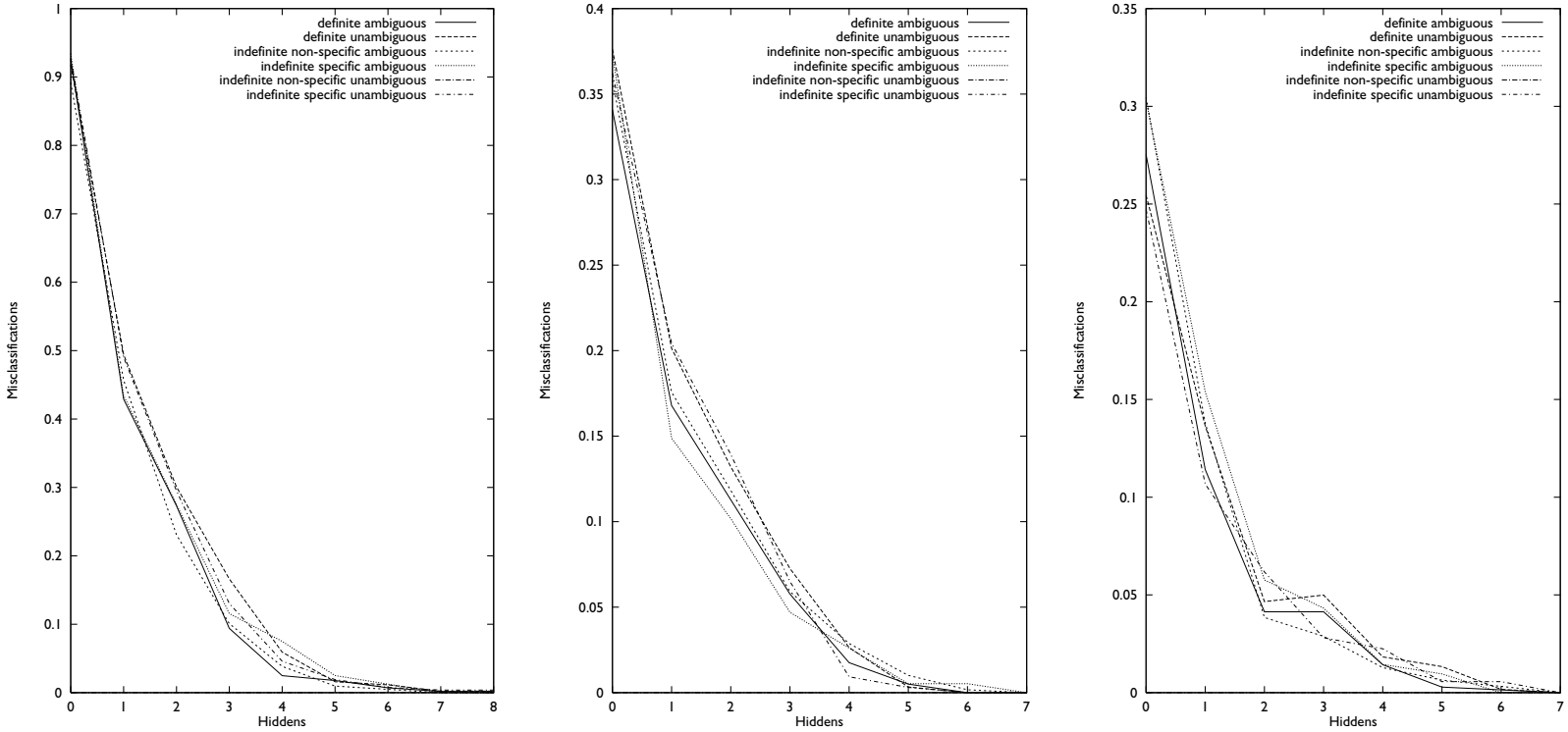
Input Epochs	Output Epochs	Average Hiddens	Average Epochs	Min/Max
150	150	3.4 (3/4)	509	(403/615)
100	100	3.3 (3/5)	444	(369/637)
100	50	3.8 (3/5)	396	(319/495)
50	100	3.6 (3/4)	466	(388/525)
50	50	3.5 (3/5)	349	(288/480)
20	20	6.1 (5/9)	257	(209/374)
10	20	defeated	–	–
10	10	defeated	–	–

(b) Configuration B

Input Epochs	Output Epochs	Average Hiddens	Average Epochs	Min/Max
150	150	7.4 (6/10)	957	(739/1230)
100	100	7.3 (6/8)	920	(789/1091)
100	50	7.8 (6/10)	796	(575/1019)
50	100	7.5 (6/9)	890	(773/1079)
50	50	7.5 (5/9)	733	(522/895)
20	20	10.4 (9/13)	436	(374/538)
10	20	defeated	–	–
10	10	defeated	–	–

(c) Configuration C

Table 5.2: Results of training averaged over 10 runs. Networks were defined as 'defeated' if they had failed to converge after recruiting 20 hidden units. †50% of trials failed. The leftmost two columns give the hard upper limits on input- and output-epochs. The figures in parentheses in both the average hiddens and average epochs columns indicate minimum and maximum values for each respectively



(a) Configuration A (b) Configuration B (c) Configuration C

Figure 5.4: Training-set misclassifications on the six situations represented in the training data for each configuration

figure 5.4(c) corresponds to the relatively large magnitude and change in sign of the weight from the 'non-specific' input to the hidden units H2, H3, and H4.

Misclassifications on novel exemplars Qualitative performance was also measured by testing the extent of the network's generalisation to novel exemplars, again measured in terms of misclassifications in each category. This data was collected using the weight matrix saved after each unit recruitment in conjunction and tested using a test set as described above with the default non-specific semantics given for the indefinite article as during training.

Figure 5.5(a) shows the test-set misclassifications over the course of training for data sets of configuration A.

5.3.3 Analysis of internal representations

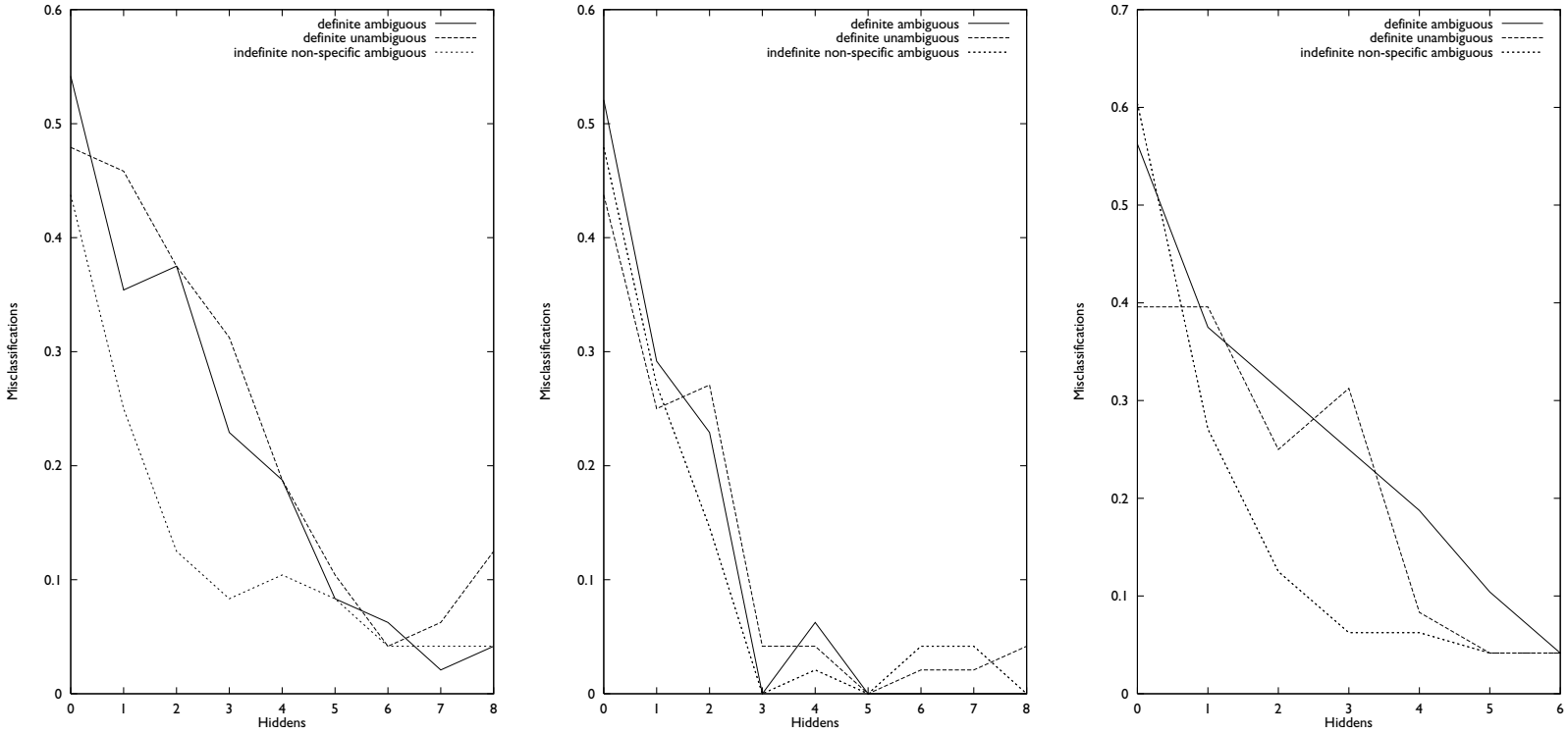
Figure 5.6 shows the final Hinton diagrams for the cases shown in figures 5.4 and 5.5. Several main features of the solutions formed by cascade-correlation are apparent from these diagrams. Firstly, relatively little attention is given to particular fine-grained features of the array itself, and where attention to these features is stronger, as in figure 5.6(a) for instance, it lessens over the course of training due to the usual effects of superposition of connectionist representations, as shown in the weights from the array inputs to the later-recruited hiddens. Each new hidden unit also attends to different patterns of question-objects.

In a network such as figure 5.6(b), very little attention is ever paid to the article itself (except by unit H4). In the other two configurations attention is initially divided more equally between the article and (all) the units in the function bank, although attention to the article itself subsequently decreases.

The sign of the weight from the 'non-specific' input also changes after every two or three recruitments. Such shifts correspond to fluctuations in the numbers of misclassifications. For instance, in configuration B, the rise in test-set misclassifications on the definite ambiguous and indefinite non-specific ambiguous categories corresponds to the recruitment of hidden unit H4 in diagram 5.6(b) which has a relatively small weight from the 'non-specific' input, but larger weights units from the definite and specific inputs than previous hidden units. In all configurations it was observed that hidden units tended to have same-signed weights from the definite and specific functions, which is what would be expected, given that both functions are used in situations where the question-object refers to a type of object of which there is only one example in that playroom.

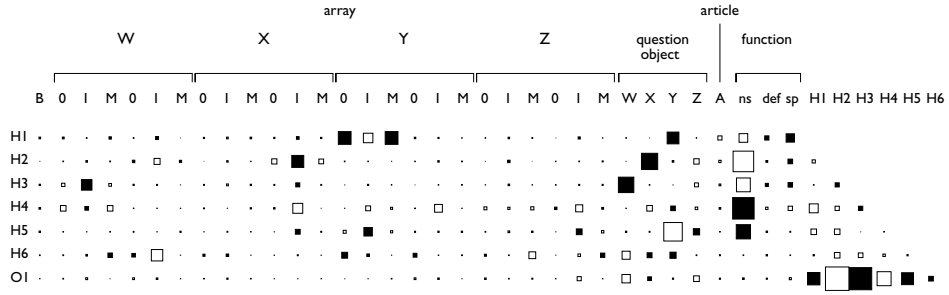
Examination of Hinton diagrams also showed that the results could be divided into a set of broad classes according to which kinds of features which the hidden units initially and subsequently focused on. These shifts in focus also corresponded to fluctuations and temporary increases in test-set misclassification rate shown in figures 5.5 as well as smaller changes on the training set. For instance the network shown in figure 5.6(c) began by recruiting hiddens which attended to the article unit and function bank with comparable strength, subsequently focusing on the non-specific function unit more strongly. This shift in attention is also apparent in the corresponding graphs of the misclassification rates on the training (figure 5.4(c)) and test sets (figure 5.5(c)) respectively. The point at which the attentional focus changes — at the point at which unit H3 is recruited — corresponds to a slight increase in training-set error on the definite unambiguous category and levelling off on the definite ambiguous category, with a decrease in error on the indefinite specific ambiguous category. In the test set, this shift corresponds to the more dramatic temporary increase in error (a micro-U-shaped curve) on the definite ambiguous category.

For configurations A and B, by contrast, the initial focus is strongly on the non-specific function unit with groups of successive recruits having same-signed weights to that unit. These shifts in sign again correspond to fluctuations in the misclassification graphs. For instance, figure 5.5(b) shows that error on the definite unambiguous class rises as the newly recruited

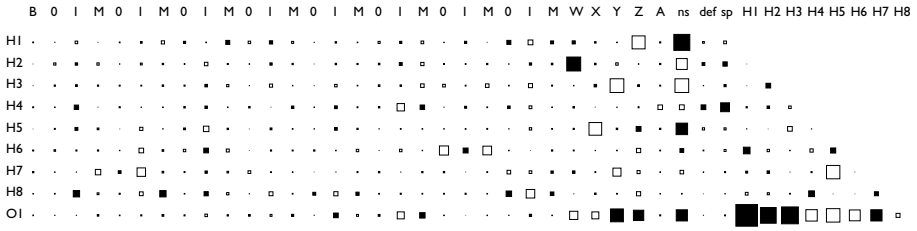


(a) Configuration A (b) Configuration B (c) Configuration C

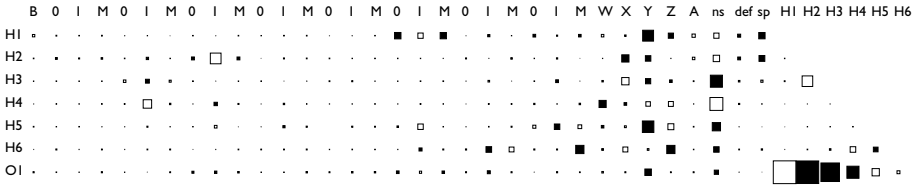
Figure 5.5: Test-set misclassifications on the definite article in ambiguous (M,1) and unambiguous (1,0) situations, and indefinite article with the (default) non-specific ('a') function and ambiguous situation.



(a) Configuration A



(b) Configuration B



(c) Configuration C

Figure 5.6: Hinton diagrams corresponding to the networks in figure 5.5. White squares correspond to negative weights, black squares to positive ones. Figure 5.6(a) provides a key to the diagrams in this chapter — each row represents the incoming weights to each successive hidden unit, with the last row representing those to the output unit. At the input, 'B' indicates the bias unit, within each object-type bank in the array the two '0 1 M' groupings correspond to the left and right playrooms respectively. Within the function bank, 'ns' and 'sp' denote the non-specific and specific senses of the indefinite article respectively. In most of the diagrams which follow, discussion will centre on the rightmost four input units, i.e., the article and function units as well as on the hidden units.

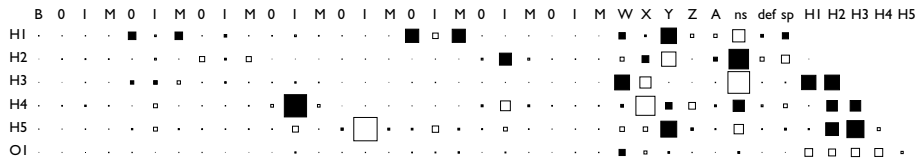


Figure 5.7: Example of the effects of training without input patience (Configuration C)

second hidden has a strong negative weight from the non-specific function unit while the first unit had a strong positive one.

5.3.4 Manipulation of internal parameters over the course of learning

As the preceding discussion suggests, configurations of category-frequency was not a strong determinant of the the patterns of shifts in focus of groups of successive hidden units which seemed to underlie the kinds of qualitative change required. In an attempt to gain more control over the nature of the feature detectors formed, as well as to investigate ways of utilising the internal resource-phasing which is characteristic of cascade-correlation, further experiments were devised using the same basic setup, but which varied internal parameters, specifically patience and pool size.

Patience

The intuitions behind varying the patience depend on the idea of overfitting and the poor generalisation which results from this. This can be linked with the ideas from the RRH of an initial phase in which representations are closely tied to perceptual features and generalisation (or systematicity) is poor.

Cascade-correlation has separate patience parameters which control input- and output-side learning. Since the weights trained by output-side learning are discarded after each recruitment phase, it was decided to focus on varying the input patience, as this has a direct impact on the weight-strength of the incoming connections to the hidden units which survive to mediate further learning.

Eliminating the effect of input patience (i.e., allowing training to continue either until victory or until the epoch-limit is reached) produced patterns of weight-strengths such as that shown in figure 5.7. The most obvious effects are the relatively strong weights from the non-specific input (compare figure 5.6(c) for instance).

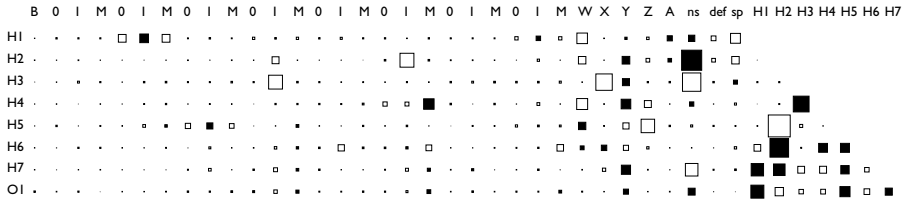
Most significant of these effects is the strong positive weights formed between particular hidden units and those recruited one and two rounds before; this indicates that earlier hidden units have a stronger mediating effect on later ones than in networks trained with low, default, patience values. Several exponential decay functions on the number of hidden units were investigated as ways of decreasing the input patience over the course of training.

Pool size

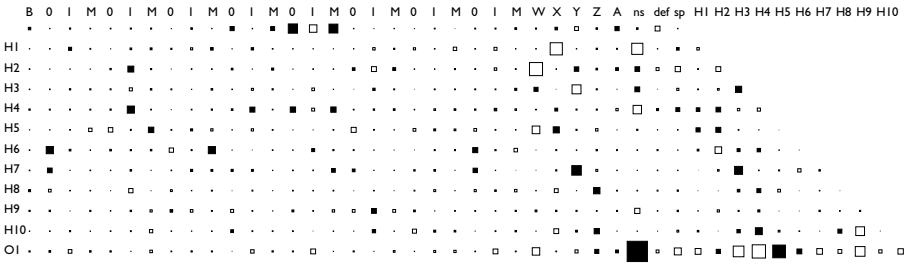
Varying the pool size controls the space of possible feature detectors which is searched for the next hidden unit to recruit. A gradient ascent search is used to find the candidate unit having the highest correlation with the error at the output units. Large values of pool-size can thus lead to sharp increases in the correlation level at each recruitment round.

The conjecture that this parameter could provide a means of controlling RR-like change proceeds from a similar intuition to that in the case of patience. A wider search space and the resulting, rather localised, optimisation of performance could lead to a kind of overfitting to the feature which is currently the most salient in the inputs. The mediation of the hierarchy of previous hiddens may even reinforce the effect over small numbers of recruitment rounds.

Training a large number of candidate units is also reminiscent of the suggestions of Bechtel (1993) concerning the production and selection of particular redescriptions of a particular task.



(a) Net trained with candidate pool of 500 units



(b) Net trained with a single candidate unit

Figure 5.8: Weights formed in networks trained with large and small candidate pools

Bechtel speculates that the process of redescription might involve the generation of a number of potential redescriptions, one or some of which are then redeployed.

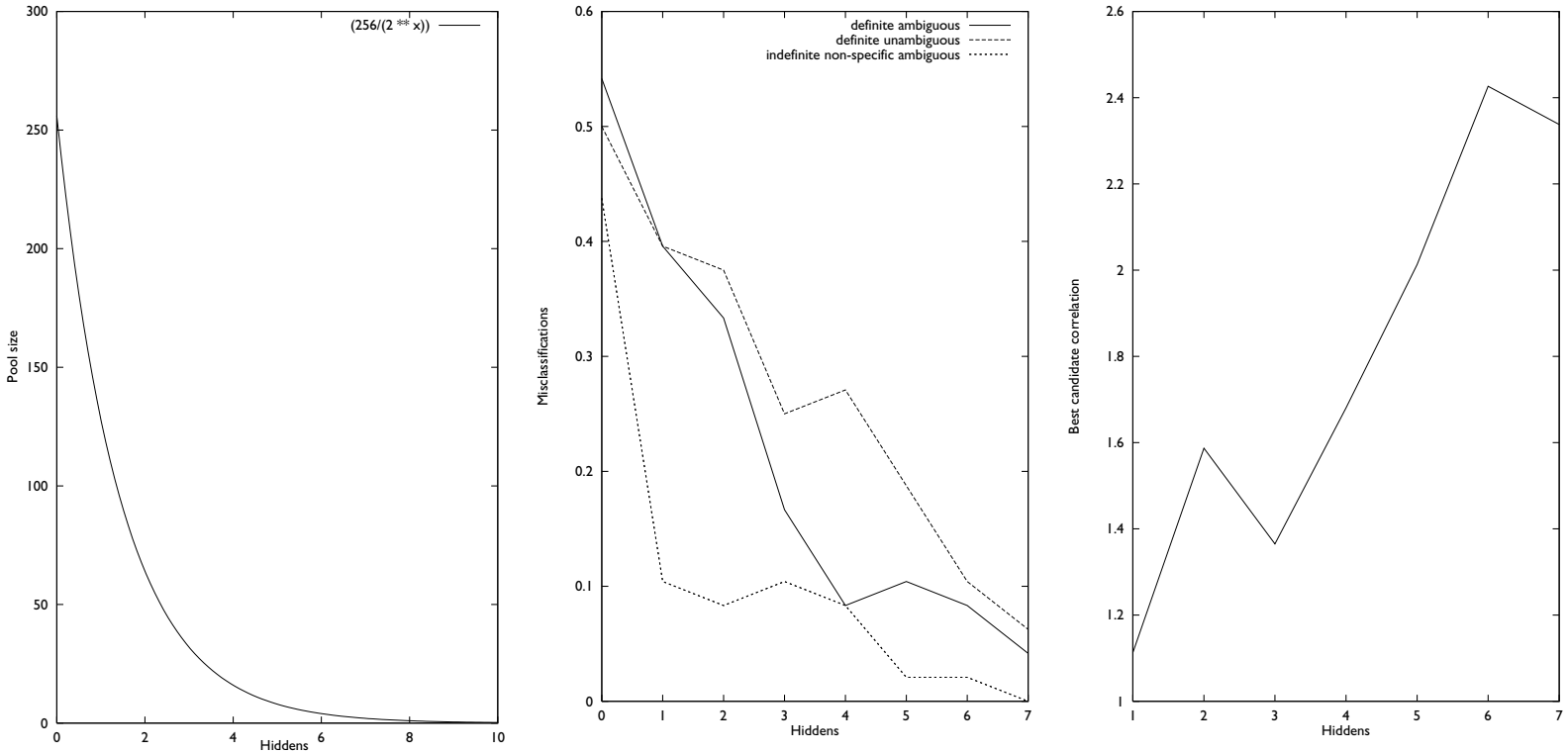
Varying pool size over the course of training The following experiments investigated the effects of different pool sizes on the same dataset, using both values which were held constant during the course of training and those which were varied according to some function.

Although there was no way to eliminate the effect of pool size completely, it was possible to examine relatively large and small values. Figures 5.8(a) and 5.8(b) show the effect on the pattern of weight-strengths of using candidate pools of 500 units (a relatively high value) and a single unit respectively. In the 500-unit case the weight pattern resembled that obtained when training without input patience.

It was conjectured that a large initial pool size would promote a kind of overfitting to the most salient features in the input. This would correspond broadly to the initial phase of RR, in which representations are disjoint and non-systematic. In order to capture the subsequent phases of RR, characterised by generalisation and the formation of feature-detectors encoding higher-level and structural aspects of the task, as well as the symptomatic U-shaped behavioural profile associated with it here, it was reasoned that the the search space should be progressively reduced by decreasing the pool size. Two schemes for reducing the pool size over the course of training were investigated.

Simple exponential decay This involved halving the number of candidates before each phase of input-side learning (i.e., at the point when the network resulting from previous training was reloaded). Figure 5.9(b) shows the pattern of test-set misclassifications and figure 5.9(a) shows how the pool size varied with the number of hidden units.

Decay given by a function of the number of hidden units As discussed in section 5.3.3 it was observed that over the course of the first few unit recruitments the network focused on input values, with weights from downstream hiddens being relatively small. After this point the network changed its focus, giving previous hiddens equal or greater weight than inputs. On the basis of this observation a pattern of pool-size decay was investigated in which the value de-



(a) Pool-size value as a function of number of hidden units

(b) Test-set misclassifications

(c) Correlation-measure *S* for best candidate unit

Figure 5.9: Effects of halving the pool size after every recruitment (configuration A)

creased gradually at first, and more rapidly towards the end of training. The following function was found to have this general profile:

$$f(h) = \begin{cases} 1, & \text{if } f(h) \leqslant 0 \\ (1 - e^{h+offset_1} + offset_2)/scaling & \text{otherwise} \end{cases}$$

where $offset_1$, $offset_2$ and $scaling$ were values needed to bring the appropriate part of the underlying graph into a suitable range for pool-size values. After a hand-search these values were taken to be 6, 20000 and 400, which compensates for the fact that the function itself takes negative values in the range of h considered (approximately 0–10 hidden units — see table 5.2). Figure 5.10(a) shows the function and figure 5.10(b) the test-set misclassifications recorded during the course of training for a typical network.

There is a drop in pool size as the point where values of the function intercepted the x-axis and the number of candidate units was taken to be a constant (here the default 8) for the rest of the training. This drop is reflected in the profile of test-set misclassifications shown in figure 5.10(b). It is these results which come closest to those obtained in the original experiment. In particular the generalisation performance on the two definite-article categories was, and remained, consistently lower than that on the indefinite category. Misclassification rates on the indefinite category also tended to rise (i.e., correct performance dropped) temporarily as in the experiment (see figure 5.2).

5.4 Further experiments: Investigating the effects of object-recognition

It is clear from the results presented in section 5.3 above that even with the simplifications to the original experiment discussed in section 5.2.1, such as reducing the number of distinct objects, the task of learning the mapping between the object information in the array and question-object banks constituted a significant part of the overall task. Two approaches were tried to investigate the effect of this subtask.

Training with more exemplars

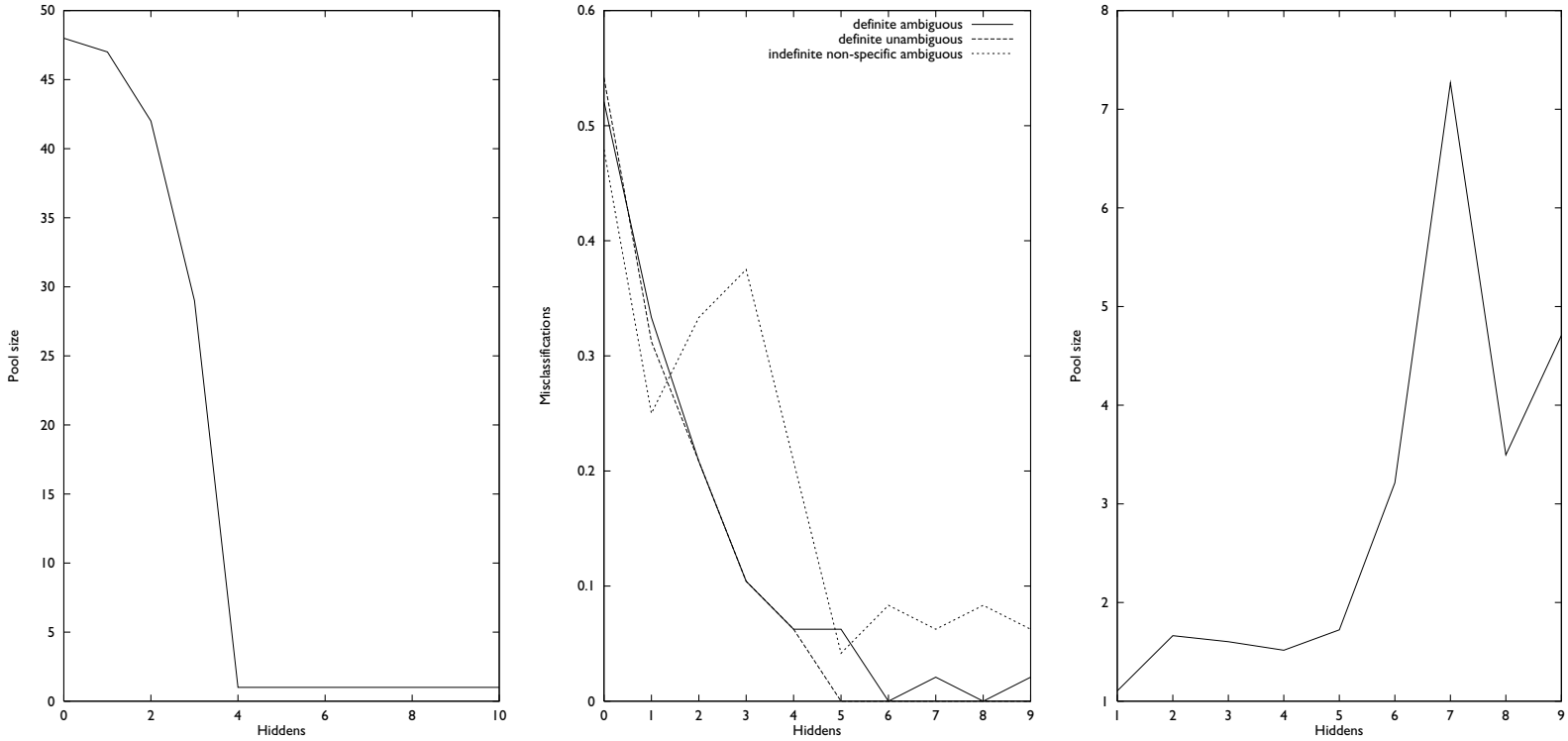
The first approach involved simply increasing the number of exemplars in the training set. A network of configuration A (i.e., 50% definite-, 50% indefinite-article exemplars) was trained with 4000 unique exemplars (twice the number used for the majority of the experiments reported above). As would be expected, increasing the number of exemplars improved both the speed of convergence on the training set and the rate at which generalisation errors decreased. However for both the training and test sets, there were far fewer fluctuations in the error profiles and the rates of test-set error on the definite article categories were consistently higher than those for the indefinite article. On these last two grounds it was decided that this approach was not worth investigating further.

Omitting the recognition component

The other variant involved omitting the recognition component of the experiment altogether, equivalent to restricting the arrays to containing objects of only one type in varying numbers. With this restriction in place, the mapping to be learned consists of just the ten associations shown in table 5.3.

In solving this simplified task the network consistently recruits a single hidden unit, doing so even when epoch limits are made very large (300 epochs) and output patience is also set at a high value (e.g., 200) to control for the possibility that in time the net could come to find a linear solution.

A typical network produced the patterns of weights shown in figure 5.11. The pattern of misclassifications on the training set was as follows. During the initial phase, (i.e., before any hidden units were recruited), the two indefinite non-specific exemplars were consistently misclassified, and with the addition of the hidden unit there were no misclassifications. Thus the basic



(a) Pool-size value as a function of number of hidden units

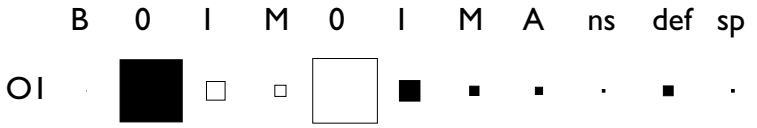
(b) Test-set misclassifications using (configuration C)

(c) Correlation-measure *S* for best candidate unit

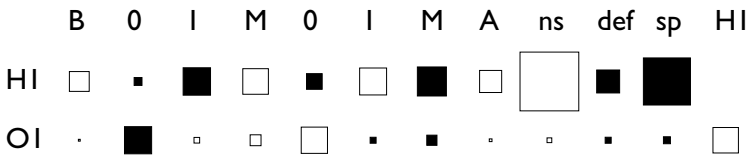
Figure 5.10: Effects of using a function of hidden-unit number to control decrease in pool size

Object	Article	Function		Addressee
(1,M)	indefinite	specific	⟼	boy
(M,1)	indefinite	specific	⟼	girl
(M,1)	indefinite	non-specific	⟼	boy
(1,M)	indefinite	non-specific	⟼	girl
(1,0)	indefinite	specific	⟼	boy
(0,1)	indefinite	specific	⟼	girl
(M,1)	definite	definite	⟼	girl
(1,M)	definite	definite	⟼	boy
(1,0)	definite	definite	⟼	boy
(0,1)	definite	definite	⟼	girl

Table 5.3: Complete mapping for playroom experiment using a single object type



(a) Before recruiting any hiddens



(b) After recruiting one hidden

Figure 5.11: Hinton diagrams for network trained on single object-type task with input- and output-epoch limits of 300 and output patience 200

progression here involves a partial solution covering all but the difficult indefinite non-specific cases, which require extra representational power.

This experiment has two main implications for the experiments involving object recognition presented above. The first is that it confirms that the underlying task requires the non-linearity provided by a multi-layer network even without the object-recognition component. The second is that the fact this version of task requires only two phases of training also seems to imply that it is the object-recognition component which must provide some of the conflicting mappings which give rise to the fluctuations seen in rates of misclassification by category.

5.5 Discussion

5.5.1 Basic performance

As we have seen, a cascade-correlation network was able to learn the basic mapping, recruiting between three and ten hidden units in the version of the task requiring object-recognition, and one in the simplified version.

Network performance compared with experimental data

The aim of these experiments was to capture the U-shaped behavioural pattern on the task of learning to map articles to functions in comprehension of French. A parallel aim was to assess whether the underlying representational progression in cascade-correlation could be said to reflect the interpretation given to these experimental results as an example of the RR model.

Before the performance of cascade-correlation can be assessed there are several obvious discrepancies between the real and simulated scenarios which should be noted. Firstly, (except in the simplified case) the network had to learn the mapping between different sources of information about particular kinds of objects, whereas children were already familiar with the kinds of objects used and were thus able to ignore the redundant information in the array with relative ease. The initial period of the network training therefore does not correspond to any in figure 5.2. Another discrepancy is the fact that in most networks, errors rates on the definite article categories were higher than those in the indefinite categories, even in configuration C, which was explicitly biased towards definite exemplars in an attempt to counter this effect.

Although we should be cautious about proposing any direct mapping between number of hidden units and the age groups of children in the original study, the following basic pattern seems to appear in general. For instance with default parameter settings, figure 5.5 shows two phases — an initial period where error decreases monotonically across all categories is followed by a period where error fluctuates. The transition between these error profiles tends to come after around a third of the total number of hidden units has been recruited.

One quantitative aspect in which network error profiles differed greatly from those obtained from children is the extent to which error rates rise during the U-shaped behavioural on the indefinite article. Even in the cases where parameter variation gave a similar overall profile to that in figure 5.2, fluctuations in network error were never as great as the rapid increase from 10% to 70% observed there, and in general network errors remained below 50% throughout.

5.5.2 The RR Model

The RR account of the comprehension performance on the playroom experiment (see section 5.1.3) conforms to the overall pattern of the RR model: in this case the phases are I, E1, at which the symptomatic comprehension errors appear, and E3, since we saw above that the children involved came to be able to express verbally their knowledge of the principles at work. Assessing such verbalisability is considered to be beyond the scope of these experiments. Discussion therefore focuses on levels I and E1, and on the overall dynamics of the whole progression.

The main innate constraints on the model are the cascade-correlation architecture itself and its initial restriction to a linear network, and the proportions of different categories of exemplars in the training set, which was controlled for by using three different configurations. In the parameter-variation experiments, the initial high values of pool size or patience can also be seen as acting as constraints on early performance and learning.

Timing of redescription

It was observed that the number of misclassified patterns in each of the three main situations considered (definite ambiguous/non-ambiguous, and indefinite non-specific ambiguous) tended to fluctuate over the course of several unit recruitments.

These effects seem to be similar to those reported by Plunkett and Marchman (1993) in that performance on inputs of similar type exhibits micro-U-shaped curves due to conflicts in mappings, here that between the form of the indefinite article in French and its two functions.

In order to interpret this kind of performance as evidence for redescriptive effects we need to consider the timing of particular qualitative shifts in behaviour in the context of those set out by the RRH.

Level I: Behavioural mastery

The RRH states that behavioural mastery of all or part of a task is a prerequisite for representational change through redescription. In the case of the playroom simulation, in terms of performance on the training set alone, since cascade-correlation is a supervised learning scheme, (criterial) mastery would seem to be the stopping condition.

Karmiloff-Smith allows that partial task mastery can lead to redescription in some domains, and here, for instance, generalisation performance on particular categories of exemplars is seen to improve before others — in particular the definite article is grasped before the indefinite. In some sense then, the network exhibits a partial mastery of the task before some stages of non-error-driven learning take place.

There are however several important differences between the timing of mastery in the network and in the children in the original experiment. In particular, Karmiloff-Smith (1979a) reports that the youngest children were already successful at the basic experimental task (i.e., comprehension alone), and this is something which cannot be claimed for the networks, whose partial solutions involve correct behaviour on subsets of categories — something which the simple (non-recognition) network brings out particularly clearly.

Level E1: Increase in systematicity

The systematicity of representations here is tested, as in Plunkett and Marchman (1993), through analysis of classifications on novel data, in this case, data in which the semantics of the indefinite article are not marked.

The RRH proposes the following progression in underlying representational structure: articles are initially represented as unifunctional homonyms, and the process of redescription acts to produce a unified and plurifunctional representation. In the simulations, there is less evidence for such a representational profile. Reasons for this include the fact that early feature detectors are sensitive to the frequencies of specific perceptual patterns as we would expect, particularly as the majority of the networks examined had to learn the mapping from array to question object at the same time as the article–function task. There is also the difficulty of designing an input encoding such that the same article is represented separately in both contexts as well as that of implementing the early one form–one function constraint believed to be pervasive in early language acquisition and to characterise the early, unifunctional forms here. The investigation of how cascade-correlation-specific parameters may be used to control overfitting was intended to model this early lack of generalisation with a degree of success.

Another characteristic of this phase is that attention shifts from perceptual input to focus on internal representations. In some cases the weights formed by cascade-correlation conform to this pattern. For instance in figure 5.6(b) the greatest incoming weights to the first five hidden units come from the input, while after this point, weights to the input are have smaller or comparable magnitudes to those from previously recruited hidden units. The output units also tend to pay rather more attention to the hidden units than to the inputs (see figures 5.6(a) and 5.6(c).

Level E3: Eventual reconciliation

As well as verbalisability and inter-domain accessibility of knowledge, the RR model states that at this level the conflicts between mappings which caused the behavioural errors symptomatic of E1 in some domains are now resolved. Most networks showed an eventual mastery of the entire task, although in some cases test-set error rates on particular categories actually rose at the final phase, and the micro-U-shaped curves which appeared after the initial sharp drop in error tended to persist right up until the point where criterial performance was reached.

Representational format

It is assumed here that basic connectionist properties such as the distributed representations formed at the hidden layers correspond to the initial implicit level of the RR model.

For the knowledge represented in the network to be considered to be at level E1 it must display a basic systematicity and generalisation to structurally similar cases. This was the case with

most of the networks discussed above, as the graphs of generalisation performance (figure 5.5) indicate.

Karmiloff-Smith (1992b) does not discuss whether knowledge of this task becomes accessible to processes from other domains, and thus it not possible to design experiments to assess this. Whether the knowledge reaches level E3 in terms of verbal expressibility as it eventually does in children, is considered to be outside the scope of this model in principle.

The role of cascade-correlation

Timing of qualitative change It was found to be difficult to make a precise mapping between the timing of individual hidden-unit recruitments and the shifts towards correct generalisation to novel exemplars in each of the three categories in the test set. However it was found to be possible to manipulate this using cascade-correlation parameters controlling the degree of (over)fitting to the particular situation presented by the inputs and the previously recruited hidden units at each stage. In particular, a large change in the size of the candidate pool could bring about a shift in test-set error rates.

These results seem to contradict the suggestions of Shultz (1994) that single unit-recruitments in cascade-correlation necessarily correspond to all three phases of the RR model.

Formation of feature detectors The kinds of feature detector formed early in training had a strong influence on whether the behavioural profile of the network exhibited the necessary kinds of qualitative change.

As noted in chapter 4, the representations formed by cascade-correlation are inherently conservative and hierarchical. Of the hidden units it is the units furthest downstream which have the greatest number of connections to other units upstream. Weightings from hidden units to the output unit tend to decrease with the recency of their recruitment reflecting the fact that early hiddens become feature detectors for the most salient features in the dataset — see figures 5.6(c) and 5.6(b) for instance.

Mechanisms of qualitative change Qualitative change is seen to be caused by shifts in attentional focus during learning. In cascade-correlation these shifts are frozen into the incoming weights to the hidden units and their influence thus persists to the extent that later hidden units and output units develop large weights to them. The output unit here developed such strong connections in all cases, while the strength of hidden–hidden connections varied, with a few general patterns of weight strength corresponding to differing behavioural profiles.

Cascade-correlation and backpropagation — error-driven mechanisms and qualitative change Cascade-correlation, in being an entirely supervised scheme, seems to share the capability of backpropagation to capitalise on residual error in exhibiting qualitative change. Plunkett (1993) suggests this could make backpropagation suitable for modelling such changes in RR. However there is an important difference between the way this affects backpropagation and cascade-correlation. In backpropagation the tendency to shift in focus between different mappings is what both causes qualitative behavioural change but is also identified (e.g., Fahlman (1988)) as a cause of slow learning performance as the hidden units 'herd' to try and capture the error-sources associated with each mapping. Cascade-correlation's freezing and single-unit recruitment mechanisms are designed explicitly to alleviate this problem by restricting and fixing the target mapping seen by the network at each stage. Thus in cascade-correlation the small rises in classification error observed are more likely to be due to the need to integrate the results of units downstream.

Thus cascade-correlation uses both residual error and its generative architecture in producing the characteristic micro-U-shaped curves.

Cascade-correlation as a model of micro-redescription It is suggested instead that the redescriptions it exhibits should be seen as micro-redescriptions which accumulate to produce larger-scale qualitative change.

In terms of the RR process, if we are to claim that individual unit-recruitments correspond even to micro-redescriptions then it must be possible to relate the strong mediation of the incoming signal to the candidate hiddens by the previous recruits to the idea in the RRH of the appropriation of the products of previous learning. The ideas discussed in Clark and Thornton (1993) provide a bridge between these two ideas of hierarchical knowledge representations via the notion of a series of feature detectors each of which recodes its incoming signal in terms of higher-order features.

The strength of hidden–hidden weights showed that previously recruited hidden structure had a mediating influence on new structure, while the strength of the final hidden–output weights shows that this structure is made use of by the outputs. This process of mediation can be seen as corresponding to the process by which the results of previous learning are appropriated in subsequent learning in the RRH.

5.6 Summary

This chapter has presented a model of redescriptive effects during the learning of the French article system, using cascade-correlation.

Some success was achieved in capturing the U-shaped curve exhibited by children in their mapping of the dual functions of the indefinite article (un/une).

Variation of the patience and candidate-pool-size parameters, which are specific to the constructive part of the cascade-correlation algorithm, was investigated as an internal means of directing incremental learning by controlling overfitting. This idea was apparently particularly well-suited to this scenario, whose initial period is guided by a one-form–one-function constraint.

It suggested that cascade-correlation gives rise to micro-U-shaped curves in a similar manner to backpropagation. Both are error-driven schemes and thus react to take advantage of residual error, but shifts in the behaviour of backpropagation are due to hidden-unit herding whereas in cascade-correlation (which is designed explicitly to avoid this effect) such shifts are due to the mediation of previous frozen structure.

The accessibility of the representations formed was not investigated here, as Karmiloff-Smith (1979a) gives no indication that the representations formed during the task become available for use by other processes either within or outside the domain. The following chapter focuses on a domain in which RR is associated with progressive accessibility of knowledge in human learners. Network transfer is investigated as a way of operationalising a test for this accessibility as well as for exploring possible constraints on the order in which the hierarchy of representational formats develops.

Chapter 6

Cascade correlation as a model of RR in sequence-learning domains

6.1 Introduction

This chapter reports results of two sets of experiments performed using the recurrent cascade-correlation architecture (Fahlman, 1991) in modelling sequence learning. These experiments explore a range of RR scenarios which complements the work on the article system presented in Chapter 5 in several ways. The addition of recurrence constitutes a difference in domain-general constraints on the network in the terms of Karmiloff-Smith (1992c, 1992a), although the incremental learning mechanism remains unaltered providing a basis for comparison between the two models. The use of the recurrent version of cascade-correlation is motivated by the focus on the learning of temporal sequences (see section 6.1.1).

The first set of experiments aims to investigate the ways in which redescription manifests itself in the increasing individuation and independence of the sequential context of sequence elements during counting.

An important distinction between the number domain and the article-function task is that Karmiloff-Smith (1992b) provides information on knowledge transfer within the number domain. It is thus possible to use task transfer between networks as a criterion for redescription in modelling this domain. The second set of experiments uses learning and structural transfer between regular grammars as a control for the influence of perceptual similarity on transfer in the counting domain.

6.1.1 Sequence learning and the RRH

Karmiloff-Smith (1990) identifies a subset of redescriptive effects which are observed across a range of domains involving sequence learning, e.g., learning to count, grasping musical structure, producing spoken language (Karmiloff-Smith, 1992b, p. 162), seriation (p. 163), and the production of written notations, as well as the learning of sequences of actions in general.

The sequential aspect to these tasks or domains is assumed to act as an initial constraint on the learning. For instance, in counting, Karmiloff-Smith (1992b) notes two properties which may act as potentiating constraints on learning: sense of one-to-one correspondence and sense of ordering. As in non-sequential domains, the RRH predicts that these constraints survive in some form in the mature version of the acquired knowledge. This is seen in the counting domain for these two constraints, for example, in the abstract idea of ordering and in relational operators.

Moving beyond innate constraints, the RRH posits that, over the course of learning, the underlying sequential representations which begin as procedural, uninterruptable wholes subsequently undergo a process of redescription. In these domains, the increased accessibility of the

redescribed knowledge manifests itself as an increased access to, or individuation of, elements of the sequence, while flexibility is seen in the ease with which sequences can be interrupted, reversed and added to, and the ease with which elements from one sequence may be introduced into another.

The diagnostic symptoms of such redescription thus centre around the ability to manipulate sequential information; in particular, its components.

Sequence learning and external notations

A further qualitative division between sequence-learning domains can be made according to whether external notations are involved. For instance, in her account of children's progressive abilities to interrupt and modify drawing sequences (Karmiloff-Smith, 1992b), Karmiloff-Smith observes that even at the initial stage, it was not the case that the drawing procedure was an uninterruptable whole as the RR model would predict. In part, she accounts for this lack of fit to the model thus:

> Drawing and all forms of *external* notation leave a trace. They also take far more time to execute, compared to the milliseconds of spoken language output, perception, and so forth. An interruption in an ongoing drawing leaves a trace of where the drawing was cut off, and it acts as a potent cue about where to continue.
>
> (Karmiloff-Smith, 1992b, p. 162)

and suggests that domains not involving external representations may provide better fits:

> I nonetheless remain convinced that representational change does exhibit initial sequential constraints, but that one may need to explore them in areas (such as counting, music, and spoken language) where no external notation is involved.
>
> (Karmiloff-Smith, 1992b, p. 162)

The work reported here thus concentrates on sequence-learning of this latter type, for which production is not assumed to involve the use of notations or other concrete external representations.

Granularity of change in sequence learning

As discussed in Chapter 2, Karmiloff-Smith (1992b) makes a distinction between the RR process and the RR model in terms of the overall progression through phases. The examples of sequence learning put forward by Karmiloff-Smith (1992b) such as counting and musical performance do not seem to follow the basic three-phase pattern of the RR model shown in figure 2.1 but rather exhibit an overall pattern of progressive explicitation which recurs over a number of phases.

For instance, in the mathematical domain, awareness of number is said to proceed from constraints on ordering and correspondence, through awareness of numerosity and explicit counting using some sort of external markers, to awareness of the link between counting and the cardinality of the counted set, to a grasp of relationships such as 'less-than' and 'greater-than', and so on, until the most generic concepts, such as that of '+1' are grasped. There is no discussion of a particular point at which this knowledge becomes verbalisable, but rather an emphasis on the increasing systematicity of the representations and the fact that the learner can increasingly reflect on that system. Again, in the rather briefer discussion given to knowledge of how to perform a musical piece, the talk is of an initially procedural whole becoming gradually more amenable to manipulation according to its components.

There are several contrasts between this pattern of redescriptive effects and that in the domain of spoken language, for instance. There the intermediate level is often marked by late-occurring errors and spontaneous self-repairs (i.e., a U-shaped behavioural curve is exhibited), and, in most cases, an eventual ability to reflect on the component structures is reported.

Sequence-learning domains

Counting According to Karmiloff-Smith (1992b), the progressive awareness of the number domain in children can be seen as an example of representational redescription. An initial mastery of counting, in which the sequence has become a routine, is followed by an awareness of properties such as one-to-one correspondence, ordinality and cardinality, which require manipulation of the components of (initially procedural) count sequences.

Gelman and Gallistel (1978, pp. 77–82) put forward a set of principles upon which counting is based and which they consider to be innate. Briefly, these are:

- The one–one principle. This involves the ticking off of items to be counted with distinct tags such that there is a one-to-one correspondence between ticks and items.
- The stable-order principle. The tags chosen to correspond to counted items must be deployed in a stable (repeatable) order.
- The order-irrelevance principle. The order in which counting is performed, i.e., which item receives which tag, is irrelevant.
- The abstraction principle. This states that the preceding three principles can be applied to any collection of entities.
- The cardinal principle. This says that the final tag in the series has a special significance as the cardinal number of the set of items as a whole.

Karmiloff-Smith is in general agreement with these principles, in particular the first three — referred to by Gelman and Gallistel (1978) as the 'how-to-count' principles — but is less sure that the cardinal principle in particular should be accorded innate status:

> It is possible that the principle of cardinality is not innately specified, as Gelman and Gallistel presuppose, but grows out of the coordination of simpler principles (such as stable order and one-to-one mapping) once these have become explicitly represented.
>
> (Karmiloff-Smith, 1992b, p. 103)

Within the RR framework, concepts such as cardinality and ordering relations are regarded as being implicit in, and redescribed from, earlier, sequential representations.

One of the general characteristics of redescribed sequences is that the increase in accessibility tends to proceed from the ends of the sequence inwards. Redescription of the counting sequence is also seen as conforming to this; in particular, the fact that the association between the final count word and the number representing the cardinality of a set precedes an awareness of concepts such as greater-than and less-than which also relate to the interior of the sequence (Karmiloff-Smith, 1992b, p. 104).

6.1.2 Connectionist models of sequence learning

Much work has been devoted to the learning of sequences of stimuli and behaviours using neural networks. Sequence learning tasks are to be distinguished from the problem of *sequential learning*, i.e., the problem of learning two or more tasks (which may be of any kind) in series in a connectionist network (although the latter also has obvious relevance to knowledge re-use and thus to connectionist models of RR). Although it is possible to apply non-recurrent networks to this task in some cases, many workers have used *recurrent* networks of some kind, i.e., networks whose underlying graphs contain cycles and which utilise these structures to make use of information about previous state, either synchronously as in discrete, locally recurrent network systems (e.g., Jordan (1986), Elman (1990a), Servan-Schreiber, Cleeremans, and McClelland (1991)) or asynchronously as in dynamical, fully recurrent systems (e.g., Yamauchi and Beer (1994), Omlin and Giles (1994)).

Counting

Several connectionist investigations of counting exist. Broadbent, Church, Meck, and Rakitin (1993) aim to capture particular quantitative as well as qualitative psychological effects. Wiles and Elman (1995) investigate the dynamics of the activation landscape of an abstract task requiring counting. In their study, a network was trained using the backpropagation-through-time scheme to predict the next token in strings from the context-free grammar $a^n b^n$. This is a counting task in that in order to perform correctly the network is required to count the number of successive *a*'s presented before the first *b* appears and then to use this information to predict the next input as each of each the *n b*'s is presented.

The models presented in this chapter differ from the above schemes in their use of incremental learning techniques, in particular the generative cascade-correlation architecture. Also, none of these previous models has examined network transfer in the context of number skills.

6.2 RCC as a model for the RR account of sequence learning

The RRH has it that the initial learning of sequences is subject to sequential constraints (such as ordering in particular) and that over the course of further learning and redescription these are relaxed so as to render the components of the sequence progressively more accessible to processing outside the original task. The particular way in which this occurs in sequences is that elements at the ends of the sequence become available before those in its interior.

So, under RR, sequences of actions, musical notes, number-words and so on go from being holistically represented and uninterruptable to being manipulable entities whose components are accessible to other processes and with which non-sequential concepts can be associated. This chapter examines how well a model using RCC could be said to fit this account.

The sections that follow present simulations of a simple task intended to relate number-sense to counting using explicit markers.

Network transfer as measure of accessibility

The RR model (see section 2.2) makes a link between the progressive explicitness of knowledge and the accessibility of that knowledge, first within and then outside its domain. Although techniques such as principal components analysis of hidden-unit activations or the localist feature-extraction technique of Greco and Cangelosi (1996b) might seem to provide evidence of the formation of structured internal representations, we also need some way of showing that network representations are structured such that they (and their components) are accessible to learning in other networks. This consideration leads directly to the idea of network transfer as suggested by Clark and Karmiloff-Smith (1993) and Clark (1993a).

However, positive network transfer cannot always be assumed to correspond to increased accessibility, since, if two tasks are perceptually similar, it may be possible for a network to learn the second simply by adjusting the positions of the decision hyperplanes defined by its weights. By contrast, the accessibility in the RR model is accessibility of the abstracted structure of the domain. This kind of transfer has been referred to in the connectionist literature as *adaptive generalisation* (Sharkey & Sharkey, 1993) or *structure-transforming generalisation* (Clark, 1993a). Structure-transforming generalisation is defined as involving 'the systematic adaptation of the original problem-solving capacity to fit a new kind of case.' (Clark, 1993a, p. 73).

Classifying transfer in networks Pratt (1994) makes a distinction between two general classes of network transfer. The first, the *related problem* class, involves transferring an entire network to a related problem on which it may already display correct performance. This kind of transfer is investigated by Pratt (1993) and Sharkey and Sharkey (1993). In the second class — *subnetwork → target network* — the source network constitutes a correct solution to a subset of the target problem, whose inputs are often a superset of those of the source network.

Which kind of transfer is required for a model of RR? Transfer in the first class relies on a combination of the proximity of the hyperplanes in the solution space and the strength of the

weights. The difficult task for subsequent learning is to move these into the positions required by the new solution. This kind of transfer does not seem to correspond well to the notion in the RRH of the transfer of structural information, since what constitutes a suitable transfer source here is determined directly by the perceptual structure of the domain.

Pratt (1994, p. 526) divides subnetwork transfer into a further two classes, corresponding to vertical and horizontal decompositions of the target network. Horizontal transfer divides networks between layers, while in vertical transfer, subnetworks span multiple layers.

The studies presented here were predominantly conducted as horizontal transfers, since the recruitment mechanism of cascade-correlation corresponds naturally to horizontal decomposition, although the transfer between counting and comparative cardinality can be seen as an example of vertical decomposition since the input and output representations were extended in the target task.

Quantifying transfer in networks There are several approaches to measuring the success of network transfer. Where the aim of transfer is to accelerate learning, one approach is to compare the number of epochs required to learn a particular task with and without the initial biasing from transferred network structure. Sharkey and Sharkey (1993) formalise this measure as

$$\tau = \frac{(\beta - \rho)}{(\beta + \rho)}$$

where β is the number of cycles required to train on a particular task from random initial conditions and ρ is the number needed to train the same task from initial conditions pre-structured through previous training, i.e., using a previously trained net. The sign of this expression corresponds to the type of transfer which takes place — where $\beta > \rho$, τ is positive, while it is negative where $\rho > \beta$, i.e., transfer has had a detrimental effect on performance. This measure allows the extent of transfer in different networks to be compared and has also been applied more recently to simple recurrent networks (Jackson & Sharkey, 1995).

Although measures such as these can give a basic impression of the extent of transfer in cascade-correlation models, the degree to which the network needs to recruit further structure and the features to which that structure attends are likely to be better indicators of positive transfer of structural information than a simple count of the number of extra epochs required. For instance, it would at least be expected that the output-side of the network would need to be retrained if the output encoding had changed, even if the structure of the knowledge were similar. Sharkey and Sharkey's scheme was thus adapted for use with cascade-correlation by calculating two values — τ_e and τ_h — which corresponded to the τ measure taken separately for epochs and hidden units respectively. While τ_e is a purely quantitative measure corresponding directly to the original τ, the value of τ_h is closer to an indicator of qualitative change as it shows the extent to which the amount of internal structure in the source network facilitates the learning of the second task.

In a short study of the role of pretraining of input–hidden (IH) weights on transfer of learning in standard backpropagation networks, Pratt (1994, p. 532) found that the best transfer results were obtained when IH weights were preset in the correct positions and weight magnitudes were raised to ensure that the hyperplanes they defined did not move out of position. The relative magnitudes of the IH and output–hidden (OH) weights were also important, but less so than the absolute magnitude of the IH weights.

These results suggest that the constructive learning mechanism of cascade-correlation already acts to preserve the half of the network which has most effect on subsequent learning. Although cascade-correlation thus embodies some of the ideas of horizontal decomposition subnetwork transfer it differs from Pratt's simple horizontal transfer networks in several ways. Firstly the IH weights of source networks were not frozen as in cascade-correlation. Secondly, in cascade-correlation, OH weights are not transferred since these weights are randomised and retrained after each phase of input-side learning. Finally, its intermediate configurations do

not capture subsets of the space of inputs as might subnetworks learning parts of a hand-decomposed problem, but rather act as a filters whose combined effect, in conjunction with the output-side weights, is to learn the task.

6.3 Counting temporal stimuli with and without explicit markers

The aim of this set of experiments was to investigate the way in which awareness of cardinality might arise through the redescription of the explicit counting sequence. It was expected that the representational change symptomatic of redescription would act ends-inwards as Karmiloff-Smith (1992b) suggests.

Since accounts of this domain within the RR framework (e.g., in particular Karmiloff-Smith (1992b)) include examples of knowledge transfer between tasks in the domain (e.g., counting knowledge becoming accessible to processes which assess cardinality), task transfer was chosen as the first indicator of redescriptive effects. We thus assume that transfer acts as a direct measure of the accessibility of the conceptualisation made by the first task to the learning of the second task.

6.3.1 Setup

The recurrent version of cascade-correlation was used to model these effects. This is essentially identical to that presented in chapter 4, except that each hidden unit has a self-connection which feeds its activation at the previous time-step back into it as an extra input. In these experiments, RCC was trained to count the number of uniform stimuli presented sequentially at a single input unit by producing the correct pattern at the output layer. The appropriate output patterns were drawn from a set of patterns representing number tags. These were encoded using a simple localist scheme which placed an inherent limit on the maximum number of items which could be counted.

Training data

Variants on marking of sequence boundaries The input used a symmetric sigmoid function, and items to be counted were represented by positive values (0.5). The end of a sequence of items was marked by resetting the state of the recurrent part of the network. This was done at the start of each count sequence as in the morse-code examples in Fahlman (1991). In most cases the end of a sequence was also marked by a negative value (–0.5) at the input unit. The output, representing the numerons (counting words), was represented using a localist one-of-n encoding.

Three training-data configurations were investigated for the counting with explicit markers tasks. These differed in the extent to which the association between count-ends and the marking of cardinality was made explicit in the input and output encodings — see figures 6.1 and 6.4. The formulation of the input and output encodings are an important source of explicit information about task structure and conceptualisation to the network, as Plunkett (1993) notes, and it was thus desirable to control for this to some extent.

Composition of training set The balanced training sets consisted of 10 sets of permutations of the set $\{1, \ldots, 5\}$. The set was permuted so that the network (which was trained in batch) would not be able to learn the task simply from the presentation-order of the sequences. These were then transformed into partial sets of sequential training data containing as many steps as the numeral, e.g., 3 in the original data set would be represented as three patterns to be presented at successive time steps in configuration A. Each permuted set thus becomes either 20 patterns in configurations B and C which included an extra input after the count sequence itself, or 15 patterns in configuration A. Fifty such sets were used in each case, giving a training set with a total of either 1000 (configurations B and C) or 750 patterns (configuration A).

There is a tradeoff between the proportions of different sequences and the proportions of different individual count words represented in any training set. For example, a training set

	A		B		C	
time	input	output	input	output	input	output
t_0	○	1	○	1	○	1
t_1	•	2	•	2	•	2
t_2	•	3	•	3	•	3
t_3	○	1	–	3	–	–
t_4	•	2	○	1	○	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 6.1: Schematic of example use of input–output configurations for the counting with explicit output markers task. • denotes a positive input representing an item to be counted, ○ an item presented simultaneously with a reset. – denotes a negative input value, or, in the output column, the output given by setting every unit to the value –0.5, a pattern which is not used to represent any numeron.

might consist of just the count sequences for the cardinal numbers from one to five. The set thus contains equal proportions of each cardinal and count sequence. However, the proportions of individual digits would be heavily biased towards the lower digits, since these are included in most sequences. In general for a set containing an equal number of each of the sequences from 1 to n, the proportion of each digit k is given by the expression $(n-(k-1))/\sum_{i=1}^{n} i$. In the setup used here, this means that while 33% of digits are 1s, only 7% are 5s. The tradeoff is that balancing the proportions of each digit represented in set would require every sequence to be 1–5. However although the relative frequencies at which children are asked to count particular numbers are not known, it seems likely that smaller numbers are inevitably more often counted through in the way that the balanced set embodies.

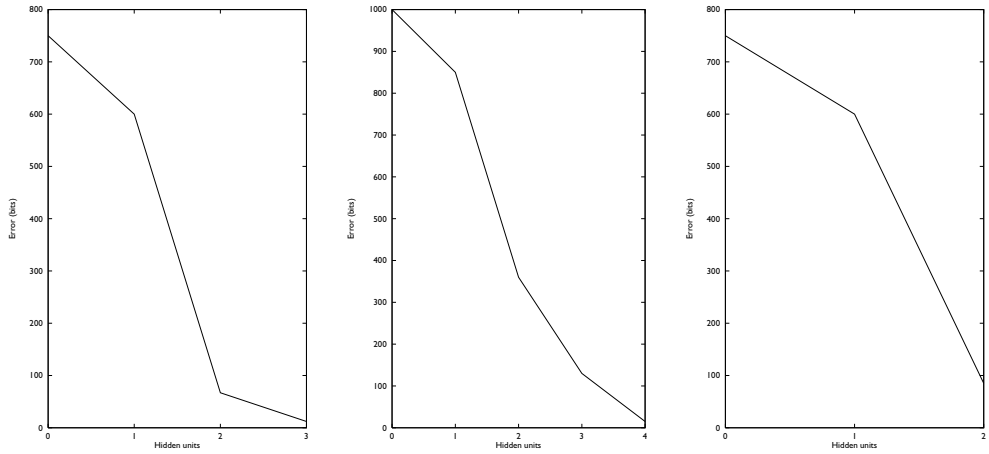
Test data

Test data consisted of input sequences of the same kind used in training. Due to the fixed one-of-n encoding used at the output layer it was not possible to use novel exemplars in the test set since the the relationships between different digits at the outputs were essentially rote-learnt. The test set consisted simply of the sequences for the number 1–5, presented in ascending order for ease of analysis.

6.3.2 Counting with explicitly marked targets

This experiment involved counting sequentially presented stimuli, using digits as explicit target-data at the output at each step. Several patterns of biasing were used in generating the training sets. Some were balanced equally between different cardinal numerals, while other sets were comprised of differing numbers of each numeral. The encoding variants for these experiments were as shown in figure 6.1.

Configuration A makes no use of input values to mark sequence boundaries and has the new count starting at the time-step immediately after the end of the previous one. Configurations B and C both have an extra step between count sequences. In configuration B, this step involves repetition at the output of the cardinal number of the count (i.e., the last number activated), while in configuration C a null (all negative) output is required. Configurations A and C both correspond to counting 'simply' (i.e., without regard to cardinal number), while configuration B reflects an intermediate state identified by Fuson (1992, 1988) in which awareness of the link between cardinality and the counting sequence is preceded by the repetition of the last element of the count (in the absence of awareness of the link between that and the cardinality of the set just counted).



(a) Configuration A (b) Configuration B (c) Configuration C

Figure 6.2: Average errors against numbers of hidden units in the counting with explicit markers task.

Performance

The basic performance on this task is shown in table 6.1. It is unsurprising that configuration C proves easiest to learn since input strobes are always associated with the same (all negative) response. Initially more surprising is the fact that networks in configuration B proved more difficult to learn than those in configuration A, even though in configuration B, sequence boundaries are indicated explicitly by a negative input. The similarity in performance between configurations A and C presumably relates to the fact that both networks are required only to perform the sequential counting aspect of the task, i.e., to activate the next right-most unit (and turn off the previously activated unit) in response to each positive input. The task in configuration B, however, also requires the network to activate the same unit as it did on the previous time-step in response to the negative input, which is clearly more difficult than the constant association made in configuration C. Figure 6.2 shows how the average number of bits of error changes with the number of hidden units for 20 networks of configuration A.

Analysis of representations

Figure 6.3 shows Hinton diagrams for networks trained on the task of counting with explicit markers in the target data. As the diagrams show, output-side weights change after each unit recruitment (this is in marked contrast with the situation in the playroom experiments reported in Chapter 5 in which the pattern and even the strength of output-side weights only changed significantly after the recruitment of the first hidden unit, i.e., when the network passed from a two-layer configuration comprising only the output-side connections to one having hidden structure).

The Hinton diagrams show that the recurrent self-connections to the hidden units all have relatively large weights associated with them and this indicates the importance of temporal or sequential information in learning the task. The first hidden unit has a large negative self-weight which causes the unit's activation to oscillate unless forced not to by the inputs. During presentation of the counting stimuli the activation of the hidden unit takes one value, switching to a value of opposing sign when the input changes at the sequence boundary. This unit thus becomes a feature detector for the only salient input feature, which is the end-of-sequence marker. Subsequent hidden units have positive weights to groupings of output units. For instance in figure 6.3, weights between H2 and output units corresponding to numbers less than three are

Input Epochs	Output Epochs	Average Hiddens	Average Epochs
200	100	3.2 (2/4)	332 (121/536)
100	100	3.3 (3/4)	334 (215/536)
100	50	3.5 (3/5)	257 (200/387)
50	50	3.4 (3/4)	246 (198/309)
20	20	3.3 (3/4)	139 (126/165)
10	10	8.0 (4/11)	172 (90/231)

(a) Configuration A

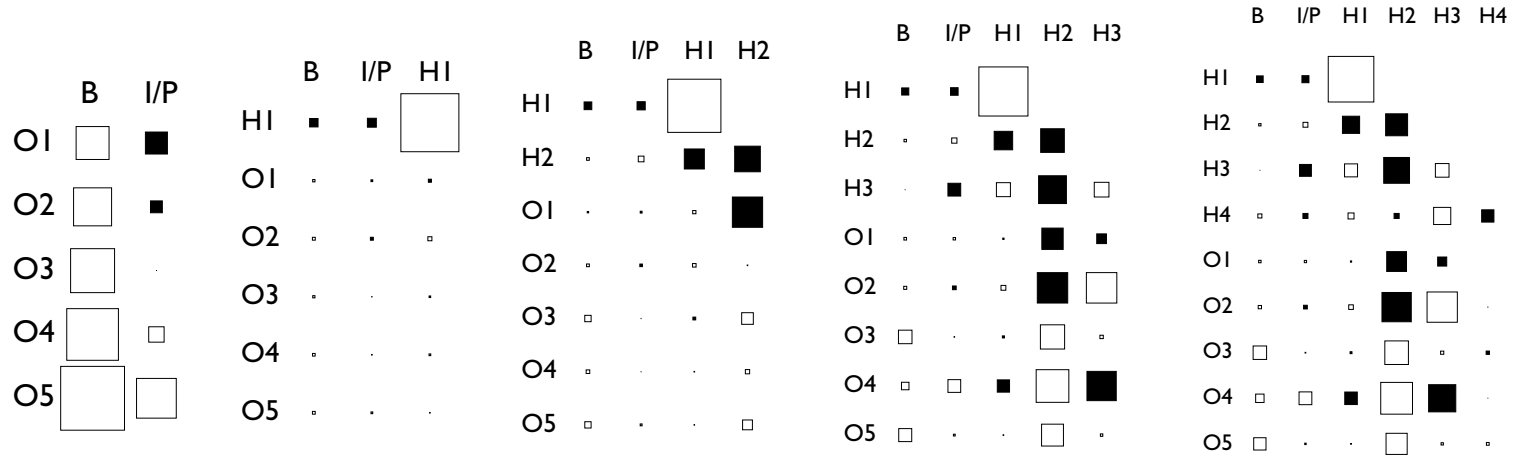
Input Epochs	Output Epochs	Average Hiddens	Average Epochs
200	200	4.5 (4/6)	505 (439/596)
200	100	4.1 (4/5)	407 (357/484)
100	100	4.1 (4/5)	434 (363/546)
100	50	4.7 (4/6)	379 (327/445)
50	50	4.5 (4/5)	346 (292/410)
20	20	6.7 (5/11)	277 (218/435)
10	10	defeated	defeated

(b) Configuration B

Input Epochs	Output Epochs	Average Hiddens	Average Epochs
200	200	3.1 (3/4)	396 (214/723)
200	100	3.0 (3/3)	291 (260/312)
100	100	3.1 (3/4)	295 (260/345)
100	50	3.2 (3/4)	246 (213/299)
50	50	3.3 (3/4)	250 (210/322)
20	20	3.6 (3/5)	152 (131/205)
10	10	11.1 (5/20)	237 (107/430) †

(c) Configuration C

Table 6.1: Basic performance of RCC on the counting-with-markers task. The leftmost two columns refer to the hard upper limits on the number of input- and output-side epochs. Numbers of hidden units and epochs in brackets indicate maximum and minimum values observed. Networks were permitted to recruit up to 20 hidden units. †20% of trials failed



(a) No hidden units

(b) 1 hidden unit

(c) 2 hidden unit

(d) 3 hidden units

(e) 4 hidden units

Figure 6.3: Weights formed at each recruitment by a recurrent cascade-correlation network trained on the task of counting with explicitly marked targets (configuration B)

time	input	output
t_0	○	–
t_1	●	–
t_2	●	–
t_3	–	3
t_4	●	–
⋮	⋮	⋮

Figure 6.4: Schematic of example use of input–output configuration (B/C) for the counting without explicit output markers task.

strongly positive, while those to outputs corresponding to higher numbers are strongly negative.

Weights on direct input–output connections are relatively small, which is what would be expected since input stimuli are identical within sequences and it is only their position in the sequence which carries any information. It is also common for the recurrent weights on the hidden units to alternate in sign (as shown in figure 6.3) or magnitude from one recruitment phase to the next.

6.3.3 Counting without explicitly marked targets

In this experiment RCC was trained to count temporal stimuli with only a final number tag as a target. The basic encoding was the same as in the previous experiment. The experiment was aimed at investigating whether an RCC network could learn a mapping between sets of temporally presented stimuli and a representation of the cardinal number of that set, presented only at the final time-step.

In contrast with the counting experiments there was only one practicable schedule for marking of sequence ends. which was to indicate the end of sequence at the input and to have the network output the total number of stimuli at the next step. The equivalent of Configuration A, in which sequence ends are indicated only by a network reset could not be used in this case since in order to output the count concurrently with the presentation of the final stimulus, the network would need to anticipate the end of sequence. The first configuration was thus used throughout, and is referred to as Configuration 'B/C' for comparison with the two configurations using similar timings in the counting with markers task (see figure 6.1). Figure 6.4 shows this configuration.

Performance

This task is clearly more difficult than the version in which intermediate targets are marked explicitly, as the network is forced to keep track of serial position internally, without any external prompts.

As in the counting with markers experiments, all the networks in these experiments had an output layer consisting of five units. Pilot studies showed that the number of hidden units needed to solve larger versions of the task increased dramatically with the highest cardinal number used, with networks unable to learn a 20-output-unit version of the task even using higher than usual hard epoch limits and parameter settings. Table 6.2 shows the basic performance on this task.

As the table shows, the performance was at its best when input and output epoch limits were set at larger equal values, e.g., 100 or 200 epochs. This suggests that input- and output-side structure are of comparable importance in solving this task (in the experiments with explicitly marked targets best performance was obtained when the input-epoch limit was greater than the output-epoch limit, in particular when these values were 200 and 100).

The profile of error in bits against number of hidden units for the counting without explicit

Input Epochs	Output Epochs	Average Hiddens	Average Epochs
200	200	3.7 (3/5)	362 (198/605)
200	100	4.4 (3/6)	325 (227/440)
100	100	5.3 (3/7)	387 (222/597)
100	50	5.1 (3/7)	332 (251/462)
50	50	5.0 (3/9)	369 (204/578)
20	20	6.0 (4/10)	247 (166/420) †
10	10	defeated	defeated

Table 6.2: Basic performance on the counting without explicit markers task, with end of count marked in input data as well as by network reset. Networks were considered to have failed if they had not converged after recruiting 10 hidden units. †10% of trials failed.

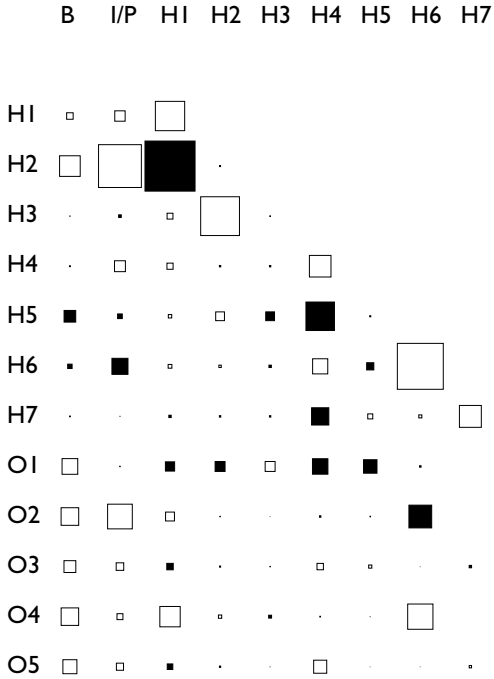


Figure 6.5: Hinton diagram of network converged on the counting without explicit markers task

markers task with marked resets (configuration B/C) is identical to that for configuration C in the counting with marked targets task (figure 6.2(c)). This shows that the recruitment of the first hidden unit makes the most difference to the error, which is to be expected, since the recruitment mechanism of cascade-correlation focuses on the largest sources of error first.

Analysis of representations

Figure 6.5 shows a Hinton diagram of a typical network which had converged on this version of the task. As the diagram shows, the overall pattern of weight magnitudes was similar to that for the final network in the counting with markers task — in particular, direct input–output connections are not heavily weighted especially relative to hidden–output connections. The first hidden unit also had a large negative recurrent weight since the input representations were the same in both setups. However, on this task networks were less likely to develop large self-recurrent weights on all of their hidden units (compare figure 6.3).

6.3.4 Transfer from counting with explicit markers to counting without explicit markers

In these experiments the weights from successful runs of the explicitly marked counting task were used as a starting point for learning on the non-explicitly marked counting. The RRH

Epochs from random start (β_e)	Hidden units from random start (β_h)	Epochs after transfer (ρ_e)	Hidden units after transfer (ρ_h)	τ_e	τ_h
		Source: configuration A			
369.0	3.5	84.25 (200/100)	0.42	0.63	0.79
		75.42 (100/100)	0.42	0.67	0.79
		92.84 (50/50)	0.67	0.60	0.68
		97.75 (20/20)	1.50	0.58	0.40
		Source: configuration B			
369.0	3.5	19.50 (200/200)	0.00	0.90	1.00
		19.50 (100/100)	0.00	0.90	1.00
		29.59 (50/50)	0.25	0.85	0.87
		33.67 (20/20)	0.25	0.83	0.87
		Source: configuration C			
369.0	3.5	137.33 (200/200)	2.17	0.46	0.24
		163.67 (100/100)	1.50	0.39	0.40
		140.59 (50/50)	1.67	0.45	0.36
		108.99 (20/20)	2.00	0.54	0.27

Table 6.3: Measures of extent of transfer for cardinal-counting networks trained from scratch, and using the saved weights from a network trained on the counting with explicit markers task

has it that redescription should make the representations formed during counting accessible to the processes involved assessing cardinality, and we would thus predict positive transfer between these tasks.

Setup

Saved weights from networks which had converged on the explicitly marked counting task were used to provide the starting point for training on the cardinal counting task. Since the input and output representations were the same for both tasks, the same training sets could simply be used to train the networks on the transfer task.

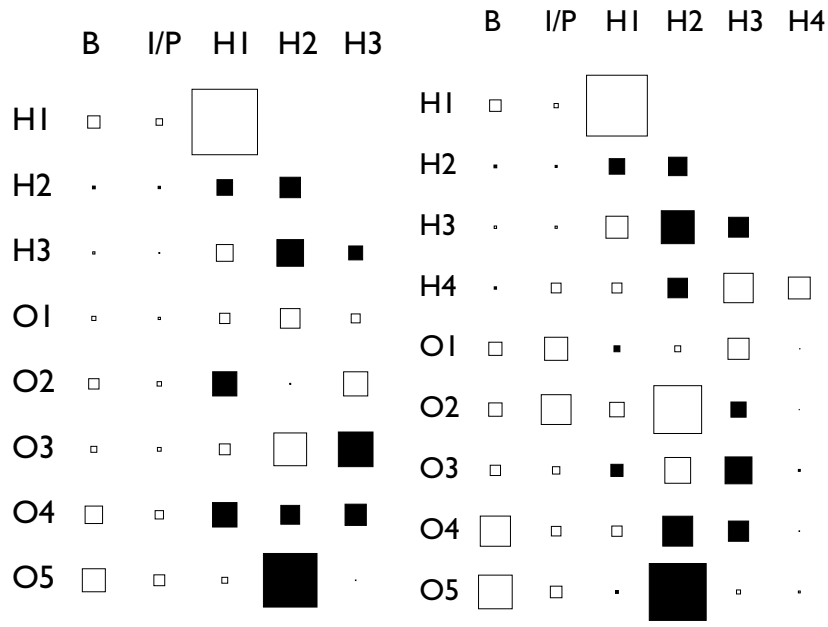
Performance of target networks Table 6.3 shows the basic performance characteristics for the transfer between explicit and cardinal counting. For these experiments, source networks were chosen which had recruited relatively small numbers of hidden units (see table 6.1 for average values).

The intuition behind this choice was that since the proportion of the total variation accounted for by each successive hidden unit decreases over the course of learning, the representations in these networks were less likely to be overfitted to the particular source data-set used to train them. Later-recruited units also attend to relatively unimportant aspects of the problem.

Relationship of representations formed in the target and source networks As noted in section 6.1 above, transfer in cascade-correlation can be measured both in terms of the number of new hidden units recruited during training after transfer (and their relationship to previously recruited structure), as well as the number of extra epochs needed to train the task.

Figure 6.6 shows the effect of transfer on network representations. Figure 6.6(a) shows the source network whose weights are used to bias training in the target network, figure 6.6(b) the weights in the target network after the completion of training on the transfer task. Pilot studies had also shown that the effect of minor variations in the number of hidden units in the source network did not significantly affect results in this case.

Figure 6.6(b) shows that the hidden unit recruited after transfer has a negative recurrent weight, indicating that it reinforces situations in which the input changes from one step to the



(a) Source network trained on counting with explicit targets (configuration A)

(b) After training on transfer task of cardinal counting (without explicitly marked targets; configuration B/C)

Figure 6.6: Effect on representations of transfer from counting with explicit intermediate targets to counting without. Unit H4 in figure 6.6(b) has been added during training on the transfer task.

next, i.e., here, sequence boundaries, which are salient to the task of producing cardinal values. However, what the diagram also shows is that the magnitudes of the weights from H4 to the output units are very small in comparison to those from all the hidden units recruited during previous training. What has more impact on the network's behaviour is the other output-side weights, some of which have changed sign or magnitude. In particular, the pattern of signs or magnitudes in the weights from each input or hidden unit reflects groupings of adjacent digits. For instance, there are large negative weights from the input to the first two hidden units, while for the bias unit, the larger weights are those to the last two outputs. This suggests that the direct input–output connections are used to code for position in the case where explicit intermediate targets are not given in the input.

Transfer from cardinal counting to explicitly marked counting

As a control on the effectiveness of the transfer from explicit counting to awareness of cardinality, transfer in the reverse direction was also examined. The results are shown in table 6.4. The transfer-profiles from these experiments differed from those in the counting–cardinality runs; transfer to configuration A was the most positive, transfer to configuration C was also possible without recruitment of any further hidden structure, while transfer to configuration B tended to require the recruitment of at least one hidden unit.

Performance of target networks

Relationship of representations formed in the target and source networks Transfer in this direction was not found to play a significant role in reducing learning time on the transfer task, with the explicit counting task recruiting as many hidden units as a network trained on that task from scratch.

Epochs from random start (β_e)	Hidden units from random start (β_h)	Epochs after transfer (ρ_e)	Hidden units after transfer (ρ_h)	τ_e	τ_h
		Target: configuration A			
332.0	3.2	18.0 (200/100)	0.0	0.90	1.0
		18.0 (100/100)	0.0	0.90	1.0
		18.0 (50/50)	0.0	0.90	1.0
		18.0 (20/20)	0.0	0.90	1.0
		Target: configuration B			
407.0	4.1	408.33 (200/100)	0.67	0.00	0.72
		412.00 (100/100)	0.99	0.02	0.61
		364.33 (50/50)	0.99	0.06	0.61
		198.67 (20/20)	1.33	0.34	0.51
		Target: configuration C			
291.0	3.0	23.00 (200/100)	0.0	0.85	1.0
		23.00 (100/100)	0.0	0.85	1.0
		23.00 (50/50)	0.0	0.85	1.0
		46.67 (20/20)	1.0	0.72	0.5

Table 6.4: Extent of benefit of reverse transfer (cardinality to counting) on performance. Source networks contained 5 hidden units in each case.

6.4 Learning comparative relations on counts and quantities

In a third set of experiments, RCC was trained to capture comparative relationships between consecutively presented pairs of sequences of counted stimuli. The single binary relation greater-than ($>$)[1] was used to investigate this. The ability to capture such relationships also provides an indication of the extent to which the network has formed a representation which reflects the ordinal aspect of the counting or cardinal numbers it has learned. The RRH predicts that awareness of such relations will appear later than awareness of the link between counting and cardinality, since they require elements anywhere in a sequence to be accessible, rather than just those at the ends.

6.4.1 Setup

In these experiments an extra unit was added to both input and output layers to encode information about relations. At the input the extra unit simply encoded whether or not a response was required at that particular time step, while at the output it is used to encode whether the relation is true or false of the two preceding count sequences.

The count input data were presented as in counting configuration B (see figure 6.1) with the ends of sequences explicitly marked by a negative input stimulus. At the output, the network was required to produce a negative value on the relation unit, except when the relation unit at the output was activated when it had to produce a value corresponding to the truth value of the relation applied to the two preceding count sequences. Figure 6.7 shows the organisation of the training data in the case $3 > 2 \mapsto \mathit{true}$.

Clearly, this task requires the network to make use of information from up to twice as many previous time-steps as in the counting and cardinality experiments presented above. Preliminary studies confirmed that the network was unable to learn the task if the recurrent state was reset at the end of every sequence of counting inputs. With resets only after both sequences had been

[1] Ability to learn this relation also implies ability to capture the less-than relation since the output values are anti-correlated.

time	input		output	
	item	>	count	truth value
t_0	○	–	1	–
t_1	●	–	2	–
t_2	●	–	3	–
t_3	–	–	3	–
t_4	○	–	1	–
t_5	●	–	2	–
t_6	–	⋆	2	true
t_7	○	–	1	–
⋮	⋮	⋮	⋮	⋮

Figure 6.7: Schematic of example use of input–output configurations for the counting and relations task on the example $3 > 2$. Figure 6.1 provides a key to the symbols used. ⋆ denotes a positive input on the relation unit, indicating to the network that a true/false response is expected at the output unit.

presented the network was able to learn the task. The training set contained equal proportions of every pair of non-equal numbers between 1 and 5.

6.4.2 Results

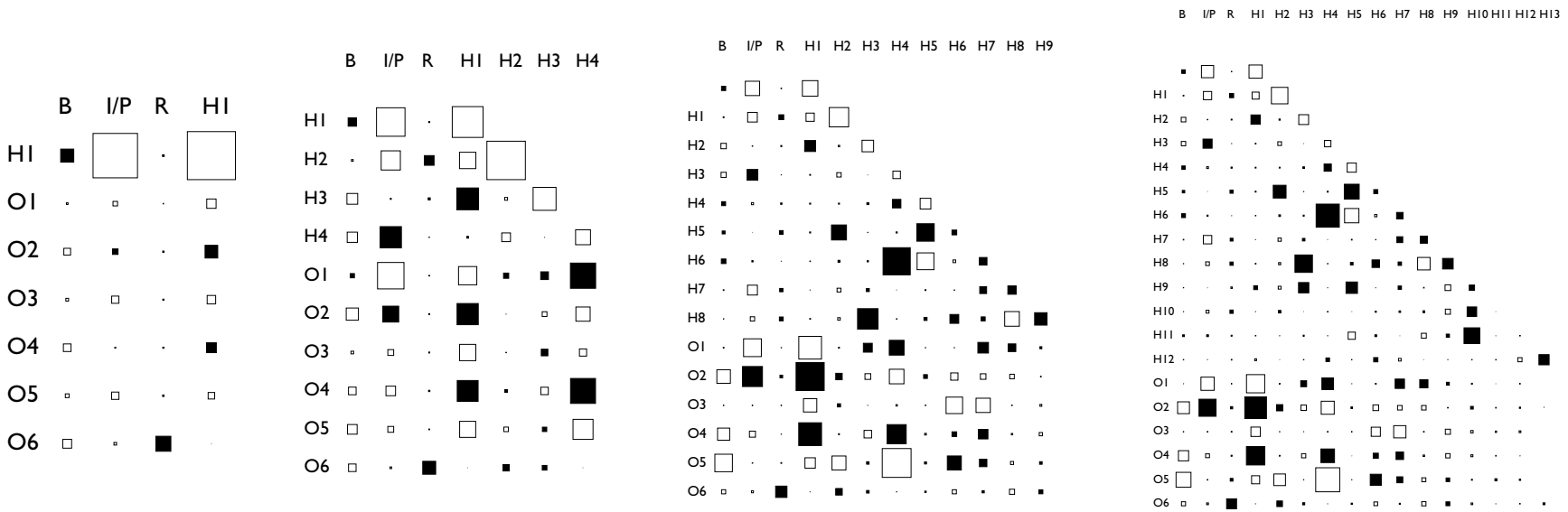
Table 6.5 shows the average epochs required to learn the task when the training set contained comparisons between all possible pairs of non-equal digits in the set. Table 6.5(a) shows that when every such comparison was included in the training set, networks required approximately twice as many hidden units to learn this task as they did in the original counting task, while, as table 6.5(b) shows, the number of units required increased when the network was required to learn the task from an incomplete training set.

Network behaviour Examination of activations at each output unit during presentation of the test set showed that the learning of the count sequence and that of the relations occurred concurrently over the course of training. Initially the activation of the truth-value output-unit corresponded directly to the end of every counting sequence (marked by input strobes). This behaviour was gradually suppressed until the unit was activated only in the case that the first sequence was longer than the second, as it should.

Analysis of representations Figure 6.8 shows the weight-values developed at intervals over the course of training on this task. As in the previous experiments, the first hidden unit acts as a feature-detector for sequence-boundaries. The first five hidden units all had relatively large negative recurrent weights indicating, as in previous tasks, attention to the change of sign at the input indicating the end of sequence. As figures 6.8(c) and 6.8(d) show, hidden units 6-10 and 13 have large positive recurrent weights indicating that they reinforce similarity in inputs at successive steps. This division indicates that the network initially forms feature detectors for sequence-ends and progressively focuses on their interiors. However, table 6.8(d) shows that hidden units 9–13 are not attended to by the outputs to anything like the same extent as previously recruited hidden units.

6.4.3 Generalisation and systematicity

In order to test the generalisation (and thus the systematicity) of the representations formed during the learning of this task, a subset of the 20 possible relations was removed from the training set for use as a test set. The test set consisted of either 8 or 4 ordered pairs of digits, half



(a) 1 hidden unit (b) 4 hidden units (c) 9 hidden units (d) 13 hidden units

Figure 6.8: Weights developed over the course of training on the relation task

Input Epochs	Output Epochs	Average Hiddens	Average Epochs
200	100	7.3 (7/8)	665 (644/688)
100	100	7.3 (7/8)	708 (657/752)
50	50	6.0 (5/8)	481 (410/616)
20	20	9.3 (8/11)	386 (328/462)

(a) Training set containing all 20 exemplars of $>$ relations on digits in the set $\{1 \ldots 5\}$

Input Epochs	Output Epochs	Average Hiddens	Average Epochs
200	100	10.0 (8/13)	1040 (875/1275)
100	100	12.0 (9/16)	994 (836/1177)
50	50	10.2 (6/17)	821 (488/1269)
20	20	15.0 (13/18)	614 (543/722)

(b) Training set of 16 relations

Table 6.5: Basic performance on the task of learning $>$-relations on different training sets

of which were positive cases, half of which were negative. Each half of the test set contained one example of each of the possible inter-number differences in order to control for these differences as indicators of the structure of the domain.

Figure 6.9 shows the training and generalisation error over the course of training. The graph shows that generalisation error is consistently higher than training-set error as would be expected, and also that it improves more quickly as the network recruits more hidden units, thus fitting it more closely to the particular exemplars in that set. However the graph also shows that it also takes comparable values throughout and eventually reaches zero which suggests that the representation of the relations formed is relatively systematic.

Initial intuitions were that relations between smaller numbers would be learned most easily, since these involved counting shorter total sequences (the total length of the sequence to be considered is given by the sum of the two numbers to be compared plus two steps at which the network is required to make a further judgement of firstly cardinality and, finally, comparative magnitude). In practice, it was found that generalisation was best on relation pairs whose difference was greatest, with classification errors (i.e., associating the wrong truth value with the inequality) only occurring in cases where the difference between the two numbers was 1. This is in some sense unsurprising since the basic mechanisms at work are quantitative. Mareschal and Shultz (1993) also reported that in their model of seriation, inputs with large differences were found to produce a qualitative improvement in performance and this reflected psychological data on that task.

It should also be possible to use hidden-unit numbers as a a further indicator of the systematicity of the representations formed during the learning of this task. For instance if the number of hidden units were close to the number of exemplars then this would imply that the network had overfitted, or effectively rote-learnt, the data. If we assume that exemplars in this case correspond to pairs of sequences to be compared, then the relationship between hidden-unit numbers and exemplars can be investigated as follows.

Data sets were generated for restricted versions of the task which used the same input and output configurations as those considered above, but which contained only greater-than relations using numbers less than highest cardinal numbers which were smaller than the usual 5. In this way, it was possible to see how the number of hidden units recruited varied with the number

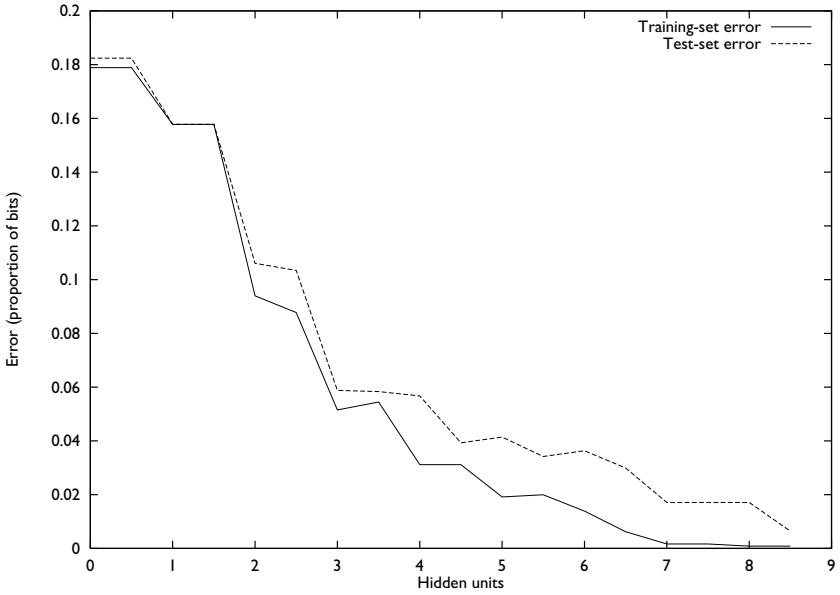


Figure 6.9: Proportional error in bits on training and test sets for $>$ task

Maximum digit	Number of $>$-relations	Lowest average hidden units (relations)	Lowest average hidden units (counting)
2	2	2.0	2.0
3	6	4.3	3.1
4	12	6.0	3.2
5	20	8.2	4.1

Table 6.6: Variation in number of hidden units recruited with increases in the number of relations to be learnt.

of relations to be learnt, and thus to assess the extent to which recurrent cascade-correlation might simply be using its ability to recruit new hidden units to rote learn the set of relations (in a similar manner to a standard backpropagation network with excessively many hidden units).

The number of relations for a given upper limit n is given by $2 \times \binom{n}{2}$, since two non-equal arguments must be chosen and these may appear in either order. The expression $\binom{n}{2}$ is given by $n(n-1)/2!$, so the whole expression simplifies to $n(n-1)$. Table 6.6 shows how the amount of hidden structure varies with the number of relations to be learnt.

Again, assuming the correspondence between final comparisons and exemplars, these figures, along with the results obtained on the within-task generalisation above, suggest that the network is evidently not simply recruiting sufficient hidden units to rote-learn the set of relations. However, comparison with the number of hidden units needed for the counting-with-markers task shows that by maximum cardinalities of 4 or 5, the number of hidden units needed for the relational task is around double that needed for counting. The next section further examines the relationship between the counting and relations tasks through assessing network transfer.

6.5 Transfer from counting to relations

Although the networks were able to learn the comparison of cardinalities task, it seems unlikely that children learn this task before they are able to produce simple counting sequences or assess

Epochs from random start (β_e)	Hidden units from random start (β_h)	Epochs after transfer (ρ_e)	Hidden units after transfer (ρ_h)	τ_e	τ_h
		Source: configuration A			
481.0	6.0	677.17 (200/100)	5.34	-0.17	0.06
		1461.67 (100/100)	5.67	-0.50	0.03
		597.50 (50/50)	5.17	-0.11	0.07
		880.67 (20/20)	15.17	-0.29	-0.43
		Source: configuration B			
481.0	6.0	873.67 (200/100)	5.00	-0.29	0.09
		799.84 (100/100)	4.67	-0.25	0.13
		606.17 (50/50)	4.34	-0.12	0.16
		552.17 (20/20)	9.17	-0.07	-0.02
		Source: configuration C			
481.0	6.0	876.50 (200/100)	5.50	-0.29	0.04
		852.83 (100/100)	5.84	-0.28	0.01
		655.34 (50/50)	5.33	-0.15	0.06
		718.00 (20/20)	11.83	-0.20	-0.33

Table 6.7: Transfer from counting with explicitly marked targets to $>$-relation

the cardinality of individual sets. The following transfer experiments were therefore designed to investigate whether prior training on counting or cardinality tasks facilitated the learning of the comparison task, and thus to what extent the representations of order and quantity formed during the original training were accessible to further learning on related concepts.

For this experiment, networks with the same extended input and output configuration as the relation networks were first trained on the counting with explicit markers task. As we would expect, results on this task were found to be essentially similar to those in the original version, since the extra units were held at constant values and thus conveyed no extra information to the network.

Table 6.7 shows the basic results for the transfer from these counting with markers networks to those learning relations. Unlike the counting–cardinality results, in this case although τ_e is negative, τ_h is positive in the majority of cases. This means that on average although the network requires more training epochs to learn the task after transfer than it would given a random starting point, transfer reduces the amount of hidden structure needed to capture the task. This result is unsurprising in some ways since marked counting can be seen as a major subtask of the $>$ task. The positive transfer between these tasks suggests that information about the cardinality of counts is represented during learning of the counting with markers task, and indeed, transfer from configuration B, in which the network was required to repeat the last count word, was slightly more positive than transfer from the other two configurations in which cardinality was not explicitly re-marked.

6.6 Transfer from cardinality to relations

Table 6.8 shows the basic performance characteristics for the transfer from the counting without explicit markers task to the comparison task. This was the only case in which both τ_e and τ_h took negative values, indicating that new hidden-unit structure was needed to 'compensate' the target network for having started from a network trained on the source task. The difficulty experienced by networks in this transfer task is in some ways unsurprising since, as the results in table 6.4 show, the trained cardinal network weights did not seem to facilitate learning of

Epochs from random start (β_e)	Hidden units from random start (β_h)	Epochs after transfer (ρ_e)	Hidden units after transfer (ρ_h)	τ_e	τ_h
		Source: configuration B/C			
481.0	6.0	880.17 (200/100)	5.50	-0.29	0.04
		1024.84 (100/100)	7.50	-0.36	-0.11
		899.67 (50/50)	7.17	-0.30	-0.09
		682.34 (20/20)	11.50	-0.17	-0.31

Table 6.8: Transfer from counting without explicitly marked targets to >-relation

Epochs from random start (β_e)	Hidden units from random start (β_h)	Epochs after transfer (ρ_e)	Hidden units after transfer (ρ_h)	τ_e	τ_h
		Target: configuration A			
262.0	3.0	19.50 (200/100)	0.0	0.86	1.00
		19.50 (100/100)	0.0	0.86	1.00
		19.50 (50/50)	0.0	0.86	1.00
		28.00 (20/20)	0.0	0.81	0.71
		Target: configuration B			
262.0	3.0	150.67 (200/100)	1.17	0.30	0.44
		141.00 (100/100)	0.84	0.30	0.56
		152.33 (50/50)	1.67	0.26	0.29
		57.17 (20/20)	1.00	0.64	0.50
		Target: configuration C			
262.0	3.0	98.50 (200/100)	1.00	0.45	0.50
		91.67 (100/100)	1.00	0.48	0.50
		111.67 (50/50)	1.17	0.40	0.44
		31.00 (20/20)	0.50	0.79	0.71

Table 6.9: Transfer from >-relation to counting with explicit markers

explicitly marked counting, and this constituted a large part of the target task.

6.7 Transfer from relations to counting and cardinality

For completeness, the reverse transfers from the relation task to counting and cardinality were also performed. Tables 6.9 and 6.10 show the results for the transfers to counting and cardinality respectively. As the table shows, τ_e and τ_h are positive in both cases.

This may initially seem surprising in the light of the predictions made by the RRH, which would predict that ability to make comparisons between counted quantities would require initial mastery of counting. However we also need to take into account the fact that the comparison network must learn to count in order to succeed when trained from a random starting point. Thus information is available which is relevant to the transfer task of counting with markers.

Epochs from random start (β_e)	Hidden units from random start (β_h)	Epochs after transfer (ρ_e)	Hidden units after transfer (ρ_h)	τ_e	τ_h
		Target: configuration B/C			
239.0	3.2	46.84 (200/100)	0.50	0.67	0.73
		46.00 (100/100)	0.50	0.68	0.73
		49.00 (50/50)	0.50	0.66	0.73
		129.17 (20/20)	1.17	0.30	0.47

Table 6.10: Transfer from >-relation to counting without explicit markers

6.8 Discussion

6.8.1 Comparison between performance of RCC and the RR account

Innate constraints

As Karmiloff-Smith (1992a) points out (and as discussed in section 3.4.1), the recurrent architecture can be seen as constituting a weak domain-general constraint on network learning. It also inherently provides the model with the more domain-specific constraints of one-to-one correspondence between items and count-terms. The design of the input encoding also enforces constraints of item- and order-indifference (Gelman & Gallistel, 1978), (although the fact that all items to be counted are identical differs slightly from the idea of item irrelevance, which suggests abstraction of the numerical properties away from a possibly heterogenous set of items).

But it is also possible that discrete recurrent network architectures embody rather too many constraints. For instance, it could be argued that a recurrent network already embodies the concept of a generic '+1' operator, and indeed Wiles and Bloesch (1992) compare discrete recurrent networks to such 'curried', or partially applied, functions.

Another concern is the discreteness of the steps themselves. Although the stepping seems to guarantee that the one–one principle holds, it is more difficult to see how the process of individuation of components in a sequence predicted by the RRH could be captured by a network which begins with components already intrinsically isolated and independent.

However, it should be noted that the previous connectionist models of counting discussed in section 6.1.2 above all used discrete recurrent networks (Broadbent et al., 1993; Wiles & Elman, 1995). Although Wickelgren (1993) also proposed the use of discrete recurrent networks, these were not restricted to being locally recurrent and thus the maximum representable cardinality was not limited as in the present study.

Subsequent learning

As noted in section 6.1.1, the pattern of explicitation in the number domain is not presented in the form of the three-phase RR model. The discussion here will thus be divided instead according to the three tasks investigated since these are assumed to require increasingly explicit representations.

Counting (with explicit markers) Networks were able to learn this task for cardinal numbers in the range 1–5, recruiting between 2 and 6 hidden units (when hard-epoch limits were set at above 50 epochs). Training sets of configuration B required a higher number of hidden units (at least 4) than those of the other two configurations. This configuration required the network to respond to the end-of-sequence signal at the input by repeating the cardinal digit at the next output step, and this required more information than simply restarting the sequence for the next count as in the other two configurations.

The RRH suggests that initial mastery at counting is underlain by a procedural representation. This is run off as a sequence which can neither be interrupted nor otherwise manipulated according to its components. In the case of the network, there was a basic sense in which the

count sequence was not interruptable in that it would be impossible to ask the network to count from any point but the beginning, and this was partly due to the fact that stimuli were identical and presented temporally. There was thus no direct way to provoke the output '3', say, without presenting three counting stimuli.

From counting to cardinality (counting without explicit markers) The RRH predicts that awareness of the cardinality of a set arises from redescription of the previously mastered counting procedure, specifically through the increased accessibility of the final element of the count sequence. In a network model we would thus expect positive transfer from counting with to counting without explicitly marked intermediate targets. As tables 6.10 and 6.3 show, transfer was positive for all source configurations. The most positive was configuration B, in which final the count was repeated, followed by configuration A, which required no response apart from the basic count sequence. This results lend support to the suggestion above that the required repetition of the final count tag requires some information about cardinality to be deployed. Fuson (1988) also identifies repetition of the last count word as a stage in the progression from rote counting to awareness of cardinality. Although it is the external emphasis on the last token which makes the cardinal number in configuration B more salient and thus a better source for transfer to the cardinal task, it should also be noted that transfer was positive in the other cases also.

Comparisons between count sequences Performance at this task was perhaps surprisingly good, considering that a locally (limited-memory) recurrent architecture was used and the network needed to deal with sequences which were over twice as long as those in the previous two experiments. In part this is accounted for by the fact that cascade-correlation is able to recruit an amount of hidden structure proportional to the number of digits involved (see table 6.6).

Since this task involves relationships between cardinal values which may be as much as four steps apart, it would seem that access to more than just the representation of the final elements is required in order to succeed at this task. As table 6.6 shows, the network did not learn the task simply by recruiting enough hidden units to represent all the possible relation-pairs explicitly, although the number of hidden units required did increase with the maximum digit used.

The RRH predicts that the accessibility of sequence elements proceeds ends-inwards. Thus in this case the development of representations underlying cardinality would precede that from comparisons, since cardinality involves only the final element. In the network model we might thus expect transfer from counting and cardinality to comparisons to be positive, and for the latter to be more positive since awareness of cardinalities would seem to be necessary for success on the comparative task. However, as figure 6.10 shows, transfer between cardinality networks and comparative networks is actually the least successful of the transfers, while transfer from counting to cardinality is positive in terms of structure (measure τ_h), but negative in terms of training time (measure τ_e).

As suggested in the analysis of section 6.6 the negativity of the cardinal–comparative transfer was due to the fact that explicit counting was a subtask of the comparative task and previous training on the cardinality task did not particularly facilitate this. This result also has implications for the accessibility of the representation of cardinality to the comparison task since, although the latter must learn the explicit counting part of that task, it should also be able to appropriate the representation of cardinality from the cardinal network to some extent, rather than being hindered by it. The result might also be taken to imply that the mechanism used by the comparative network represents cardinality in a way which is not divorced from the count sequence as it is in the cardinality task.

The positivity of transfers in the reverse direction also points to the similarity between the counting and comparative tasks, as transfers from comparative networks to both counting and cardinal networks are positive. These results also imply that some of the staging of learning in the comparative task is due to the cascade-correlation architecture alone. As the analysis of the activation patterns showed, the network learned the counting and comparison subtasks concurrently and its problem decomposition thus differs from that which was hand-engineered

		Initial training		
		Counting	Cardinality	>-relation
Transfer training	Counting	–	$\tau_e +, \tau_h +$	$\tau_e +, \tau_h +$
	Cardinality	$\tau_e +, \tau_h +$	–	$\tau_e +, \tau_h +$
	>-relation	$\tau_e -, \tau_h +$	$\tau_e -, \tau_h -$	–

Figure 6.10: Overview of results of transfer between counting, cardinality and relation experiments with RCC

through network–network transfer.

Correspondence between RR and recurrent cascade-correlation Servan-Schreiber et al. (1991) and Cleeremans (1993) describe a representational and behavioural progression in simple recurrent networks, which begins with the correct treatment of individual elements, then bigrams (pairs of elements), and finally whole sequences or paths through the idealised machine. This progression seems to correspond better to the process of proceduralisation. In some ways proceduralisation is to be regarded as a complementary and opposing process to RR (Karmiloff-Smith, 1992b, p. 17), although it could also be seen as corresponding to the initial progress towards behavioural mastery in domains, such as piano playing and drawing (Karmiloff-Smith, 1992b, 1990), in which fluent sequences must be composed from initially distinct elements.

However the results presented above suggest that the RCC architecture is able to capture some of the effects on sequence learning described by the RRH. In particular the ability to learn comparisons between different cardinalities is an example of the progressive differentiation of elements inwards from the boundaries of sequences over the course of learning.

6.9 Learning structured sequences with RCC

As noted above, two of the constraints on counting put forward by Gelman and Gallistel (1978) are those of order-invariance and item-invariance, which state respectively that neither the order nor the identity of items to be counted should affect the result. The latter was embodied in the input encoding chosen for the preceding experiments. The item-identity property was enforced in the decision to make all count items identical. Thus, as a complementary study to that presented above, the behaviour of recurrent cascade-correlation during both learning and transfer was also investigated for structured sequences, defined as those in which both item identity and serial order are relevant to the task.

These effects were investigated in the context of the learning of regular grammars. Although this is not a domain which has been explicitly linked to the RRH, the finite-grammar framework has been used to investigate abstract navigation tasks (Basye, Dean, & Kaelbing, 1995; Mozer & Bachrach, 1991; Chrisley, 1993; Chrisley & Holland, 1994). One of these — the Connectionist Navigational Map (Chrisley, 1993; Chrisley & Holland, 1994) — was also devised as a setting for the investigation of the development of objectivity and systematicity in networks. Also of interest here are those studies which have used discrete, locally recurrent networks to capture the effects of implicit grammar learning in human subjects (Cleeremans, 1993; Dienes et al., 1995).

Another motivation for this study was to provide some comparison with previous work on the representations of such grammars formed in discrete recurrent networks (Servan-Schreiber et al., 1991; Cleeremans, 1993), and also on the structural transfer of such knowledge (Dienes et al., 1995; Jackson & Sharkey, 1995; Dienes et al., submitted) (although these latter studies have been conducted using simple recurrent networks (Elman, 1990a), not recurrent cascade-correlation as in the present study).

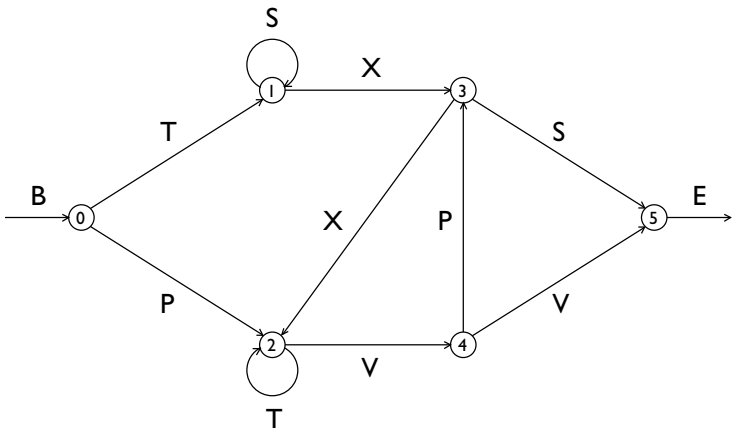


Figure 6.11: The finite-state machine accepting the Reber grammar

6.9.1 Structural transfer between isomorphic machines

This experiment was intended to assess the extent to which RCC forms representations which are independent of particular perceptual inputs. The task requires the network to transfer the ability to predict the next machine state using a particular finite-state grammar, to exemplars of another grammar which is isomorphic to the first but uses a different input language. In terms of the machine diagram associated with the grammar, the transfer is to a machine whose underlying graph has the same structure but whose arcs have been relabelled. This task is inspired by that used by Dienes et al. (1995, submitted) and can also be related to the work of Jackson and Sharkey (1995). The aim of the present experiment was to determine both whether RCC was capable of forming feature-detectors for structural rather than perceptual elements of the state-machine, and whether it did so spontaneously. This was assessed, as in the counting experiments above, by the extent to which these feature detectors could be re-used for solving tasks with the same structure but which were perceptually dissimilar.

Setup

A standard RCC network was used to learn the Reber grammar, as in Fahlman (1991) and using the parameters given by Fahlman (personal communication, 1995). The Reber grammar is shown in figure 6.11. This grammar is considered to be particularly challenging to locally recurrent networks as it contains both cycles (between states 3, 2 and 4) and star-closures[2] (on nodes 1 and 2).

The basic input encoding was as in Servan-Schreiber et al. (1991) and Cleeremans (1993), with tokens of the input alphabet {B,S,T,V,P,X} represented in a localist (one-of-n) manner at the input layer. The input 'B' indicates the beginning of a sequence and is accompanied by a reset of the recurrent part of the network as in the counting experiments reported above. At the output layer a localist encoding was used to represent the six states of the machine recognising the grammar, giving six units in each layer (excluding the input bias unit). The symmetric sigmoid function was again used, giving binary values of ± 0.5.

As in the study by Fahlman (1991), a training set of 128 unique patterns was used without any length restrictions being placed on strings (Fahlman, personal communication, 1995).

These studies differed from those of Dienes et al. (1995) and Jackson and Sharkey (1995) in that rather than requiring the network to predict the next state — a task which for grammars such as the Reber grammar is not deterministic since there are several possible next states in each case — it was required to give an output corresponding to that state. This is essentially

[2]Defined as an indefinite number of repetitions of the symbol labelling the self-connected arc.

Input Epochs	Output Epochs	Average Hiddens	Average Epochs
400	400	3.0 (2/4)	415 (265/591)
200	200	4.4 (3/6)	471 (317/599)
200	100	4.4 (3/6)	471 (317/599)
100	100	4.8 (2/6)	427 (172/591)
50	50	3.8 (3/4)	335 (257/376)
20	20	13.0 (10/18)	273 (210/378)
10	10	14.4 (10/21)	304 (210/445)

Table 6.11: Basic performance on the Reber grammar

the scheme used in Chrisley and Holland (1994), in which an agent must learn to output the description of its current location. A finite-state machine extended in this way to include a function from states to a set of output symbols is known as a *Moore machine*[3] (see Hopcroft and Ullman (1979) for instance). Basye et al. (1995) also present their model of navigation in terms of Moore machines.

This task differs from the predictive task in several ways. Although the net must still use the sequential information associated with legal successors in the grammar (since the input data remains the same), the output task is now the simpler (and non-ambiguous) one of 'responding' with the output pattern corresponding to the correct symbol of the output alphabet of the notional Moore machine induced by the network. Since only one response was legal at each step, error-tolerance measures could now be used to assess convergence.

6.9.2 Method

In the following experiments networks were trained to induce Moore machines based on the grammar recognised by the machine in figure 6.11. In the basic version of the task the output alphabet was given simply by the identity function on the state-labels in the machine diagram. These were then encoded using a one-of-n scheme.

Basic results

Table 6.11 gives the basic performance for the Moore-machine version of the Reber grammar task. On the predictive version of this task, Fahlman (1991, p. 5) found that recurrent cascade-correlation converged in all cases, after recruiting an average of 2.1 hidden units and after an average of 195.5 epochs' training. This suggests that the Moore machine version of the task is somewhat harder for the network than the predictive one, presumably since explicit, unique targets are provided at each step.

Encoding functions in data generation

In order to explore the effects of varying mappings, the work of generating training data from the abstract symbols associated with the grammar was divided between two functions:

- **Preprocess function** This maps symbols in the input alphabet, Σ, onto input patterns.
- **Output function** This maps states in the machine (considered as encoded by the input–hidden connections) to output symbols. This function is just the λ function in the corresponding Moore machine or, in the Chrisley and Holland (1994) scheme, the description function.

[3] More formally, a Moore machine is a 7-tuple $(\Sigma, q_0, Q, \delta, F, \lambda, \Delta)$, with the usual finite-state machine transition function $\delta : \Sigma \times Q \to Q$, and an output function $\lambda : Q \to \Delta$, which maps states in Q onto symbols in the output alphabet Δ.

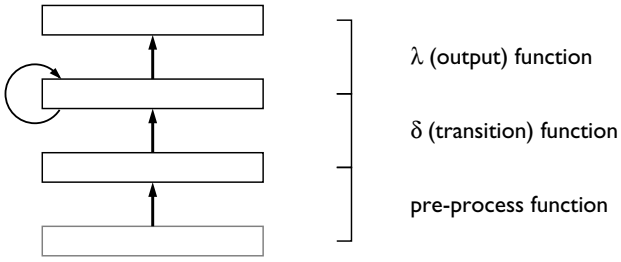


Figure 6.12: Correspondence between network resources and machine functions in recurrent networks. The central layer of those shown in black is intended to represent some mechanism of local recurrence, such as the layers of self-connected hidden units in RCC.

Epochs from random start (β_e)	Hidden units from random start (β_h)	Epochs after transfer (ρ_e)	Hidden units after transfer (ρ_h)	τ_e	τ_h
		Transfer between output functions			
173.833 (78/334)	1.733 (1/5)	52.000	0.000	0.539	1.000
		Transfer between pre-process functions			
81.933 (41/148)	2.107 (1/5)	392.444	3.056	–0.655	–0.140

Table 6.12: Extent of transfer between networks with different input- or output encodings

Figure 6.12 shows these functions in the context of a generic locally recurrent multi-layer architecture such as RCC or an SRN. This division was inspired by Dienes et al. (submitted), whose network includes a preliminary layer of what are referred to as 'mapping' weights intended to complement the encoding role of the hidden–output weights. In our models, the (external) pre-process function has the same role as mapping weights.

Transfer between encodings

To simplify the input and output encodings, sets of possible states were permuted to give transfer tasks. Transfer between isomorphic but perceptually distinct machines was investigated in two different situations. In the first, the transfer task involved outputting different output symbols for each state (formally equivalent to permuting the order of elements in the co-domain of the output function of the Moore machine). The second involved using different input symbols. This was achieved in practice by permuting the co-domain of the pre-process function which mapped input symbols to their representations as input vectors. Another way of viewing these two variants is that the first has the effect of permuting the labels on states, while the second permutes labels on transitions.

6.9.3 Results

The τ measure of Sharkey and Sharkey (1993) was again used as a way of quantifying the extent of transfer. Table 6.12 shows the results for the transfer between different input and output encodings.

In terms of structural transfer, measured by τ_h, it is obvious that while no new structure is needed to learn a new output function, transfer between networks with differently labelled transitions is consistently negative. Networks in this case recruited at least as many hidden units after transfer as before, implying that they could not make any use of the machine structure already encoded by the source network. In terms of extra training time in epochs, the values of the measure τ_e again indicate that while some retraining was necessary in the output case,

negative transfer in the input case is even more pronounced than for the structural measure.

These results are consistent with those obtained for SRN's trained on the predictive version of this task (Cleeremans, 1993; Jackson & Sharkey, 1995; Dienes et al., 1995) — new output encodings are easily learnt simply by retraining the hidden–output weights, whereas new input encodings require a new transition structure to be learnt from scratch by the recurrent input–hidden part of the network.

The reason for this difference is made clear by considering a correspondence between network resources and machine functions analogous to that made by Chrisley and Holland (1994) for the SRN's in their study. In terms of Moore machines, the hidden–output connections compute the λ (output) function while the input–hidden connections are seen as computing the δ (transition) function as in the predictive model. For recurrent cascade-correlation, the correspondence is similar, but it is the whole of the (multilayer) input-side structure which implements the transition function.

Dienes et al. (1995) successfully compensated for this asymmetry between network resources and machine functions by extending an SRN with an extra layer of weights preceding the usual input layer. The network was then able to use this extra layer to process changes in input mappings in the same way as the hidden–output weights did at the output. The results of the current study imply that the multilayer architecture of cascade-correlation does not come to act like the preliminary layer in Dienes et al.'s model (partly because of freezing), and that in this way the architecture is just as bound by perceptual cues in transfer as more conventional backpropagation-based models such as the SRN.

6.10 Summary

The effects of redescription on knowledge of sequential information appear in a variety of domains and tasks including seriation, counting and musical performance. In some cases, although the redescriptive process appears to act to give progressively more accessible representations, the three-phase pattern of the RR model is not observed. This chapter has presented a study of the use of (recurrent) cascade-correlation to model redescriptive effects in domains involving sequential data, specifically in the domain of number. The use of the locally recurrent version of cascade-correlation provided an additional domain-general architectural constraint in the way suggested by Karmiloff-Smith (1992c).

In the domain of temporal counting, networks were trained on three related tasks of increasing difficulty: counting with intermediate counts explicitly marked at the target, counting without these intermediate markers and with only the end token (representing the cardinal value) marked, and counting pairs of sequences then classifying the pairs according to whether the cardinality of the first sequence was greater than that of the second. The last of these could also be viewed as the learning of the ordinal relationships between numbers.

Karmiloff-Smith (1992b) claims that the representations formed in simple counting are made accessible to those assessing cardinality through redescription. The general properties of redescription on sequences suggest that relational properties between numbers in the middle of sequences would become accessible only after properties such as cardinality, which involve only the end-point of the count.

Accessibility of representations to processes associated with different tasks was measured directly through task transfer. It was assumed that the positive transfer of task structure provided an indication that task structure was in a suitable form for re-use by structurally related tasks. Sharkey and Sharkey (1993)'s measure of transfer was used and also applied to the amount of new structure added by cascade-correlation after transfer.

Results indicated that counting with explicit markers transferred postively to counting without such markers, particularly in the case where the last element of the count was repeated. Transfer was also positive (in terms of amount of hidden structure) between counting with markers and comparative counting. However, previous training on the cardinality task was actually

detrimental to learning on the comparisons task.

The reverse transfers from cardinality to counting and from comparisons to cardinality and counting were perhaps surprising in that previous training on comparisons facilitated learning of both counting and cardinality. This was thought to be due to the fact that the counting is a major subtask of the comparison task. The staged learning inherent in the cascade-correlation scheme evidently forms intermediate representations of the count in this task which are usable during subsequent learning on a counting task alone.

Some concerns remained about the burden placed on innate constraints in the model, in particular the fact that inputs were pre-segmented. Others of these constraints however corresponded to those put forward by Gelman and Gallistel (1978), such as one-to-one correspondence and item- and order-irrelevance.

As a control for the perceptual similarity of stimuli in the counting tasks, a complementary study was performed involving transfer of structure between sequences in which the order and identity of items was important. The Reber grammar was used in experiments to test the extent of transfer between different mappings from perceptual labels to an isomorphic structure.

These experiments were less successful, with networks exhibiting strong negative structural transfer. It seems apparent from these experiments that the input–hidden structure acts as the transition function of the state-machine, but does not also perform the re-encoding of inputs of the initial layer in the model of Dienes et al. (submitted), and is thus not able to transfer the function positively to isomorphic tasks without adding a comparable amount of new structure. These results suggest that cascade-correlation does not tend to develop representations which are suitable for structural transfer in the general case.

Chapter 7

Skeletonisation as a model of representational redescription

Introduction

In Clark and Karmiloff-Smith (1993), skeletonisation of networks was proposed as a promising technique for capturing redescriptive effects, in particular the idea that redescription results in reduced representations which preserve the relevant features of the original learning, as discussed in section 3.8.1. Despite the criticisms presented there, skeletonisation still exhibits an interesting set of incremental learning techniques which both intersect and complement those of cascade-correlation. In particular, although it is also based on an error-driven scheme (backpropagation) and involves phasing of internal structural resources, there are other features which correspond more closely with aspects of the RR model. For instance, skeletonisation explicitly acts only after the error-driven network has converged and the process of relevance assessment acts off-line and is not directly related to error during training.

7.1 Skeletonisation and RR

Skeletonisation (Mozer & Smolensky, 1989a, 1989b), is a scheme in which units are pruned from either the hidden or input layers of a network trained using backpropagation, according to their *relevance* to the reduction of the overall error.

Three claims are made for the potential usefulness of such a technique. Firstly, it is claimed that by reducing the number of hidden units the network will be constrained to produce better generalisations, secondly, since learning is fast with larger numbers of hidden units, the technique should accelerate learning by allowing the initial phase to produce many possible generalisations while the later phase constrains those generalisations, thirdly, it should be possible to gain a better understanding of a skeleton network, since although the process may result in a decline in the percentage of correct answers, it may become possible to analyse the resulting network in terms of a small number of rules.

The first and third of these claims are particularly relevant to aspects of the redescription process. As Clark and Karmiloff-Smith (1993) note, redescription seems to produce improvements in generalisation at the expense of accuracy in performance, and speculate that this is because the resulting representations are in some way 'reduced' by the process. The idea that such networks should be easier to interpret in terms of rules can also be thought of as corresponding to an increase in explicitness of the network's internal representation. Finally skeletonisation acts only after error-driven learning has succeeded, and this seems to correspond to the idea that redescription acts only after behavioural mastery has been achieved.

Mozer and Smolensky (1989b) provide several examples in support of these claims of improved learning and improved intelligibility. For instance, in a stimulus-sorting task the skeletonisation procedure correctly trims away all the inputs surplus to the network's ability to sort

stimuli into two classes. Other examples (the four-bit multiplexor and random mapping problems) show that the network is better to learn a particular mapping than a standard backpropagation network forced to work throughout its training with only the same number of hidden units as the final skeleton network. According Mozer and Smolensky this is due to the initial over-provision of resources in the skeletonised network.

7.1.1 Skeletonisation and other pruning techniques

Pruning (of connections or units) is an established technique in connectionist engineering – see Reed (1993) for a survey of pruning techniques. Mozer and Smolensky (1989b) relate relevance to other measures such as contributions (Sanger, 1989) as well as to the analysis of internal patterns of representation using cluster or principal components analysis.

An alternative approach is to use some sort of cost term related to the complexity of the network such as the size of the weights, the number of connections, hidden units or hidden layers, or the symmetries of the network .

The idea of an initial overprovision of resources followed by a pruning procedure also corresponds to ideas about cortical maturation which have recently been related to developmental phenomena. For instance Johnson and Karmiloff-Smith (1992) discuss the application of principles derived from the study of selective neuronal loss during development to aspects of cognitive development and language acquisition.

7.2 The skeletonisation procedure

7.2.1 Calculating relevance

The relevance of a particular unit is an approximation to the difference between the overall network error on a particular training set with that unit installed, and the overall error with it removed. Since computing this value for every unit would involve a separate pass through the entire training set (if indeed such a fixed training set exists) a computationally less expensive measure is used.

In the experiments which follow, expressions for relevance were derived, as described in Mozer and Smolensky (1989a), in a similar manner to standard backpropagation. As suggested there, the linear error measure

$$E^l = \sum_p \sum_j |t_{pj} - o_{pj}|,$$

is used in preference to the quadratic measure normally used in backpropagation, since the derivative of the latter tends to zero as the total error decreases.

The additional information was noted (Mozer, personal communication, April 1994) that, using the linear error measure the expression for δ_i at the output layer is given by

$$\delta_i = f'_i(net_i)\nu_i,$$

where $\nu_i = -1$ if $o_i > t_i$ and +1 otherwise. Relevances at the hidden or input layers are then given by

$$-\frac{\partial E}{\partial \alpha_j} = \left[\sum \delta_k w_{kj}\right] o_i$$

where k is an index over the units in the layer above. In this implementation, linear errors are computed at each presentation to give the δ_is at each layer.

7.2.2 Which layer to skeletonise?

It is possible to apply the skeletonisation procedure to the input or hidden layers of a backpropagation network, as discussed by Mozer and Smolensky (1989a, 1989b). It was considered inappropriate to delete units from the input layer, especially since, when using a localist input/output representation as in the data used with cascade-correlation in chapters 5 and 6, after the deletion of a number of input units, some previously distinct patterns would effectively be represented by the same (null) pattern. This would constitute changing the task to be learned rather than just the structure of the representation underlying the network's solution. Thus in the experiments that follow, skeletonisation was performed on the hidden layer in the interests of manipulating internal representations.

7.3 Choice of experimental tasks

The study presented here was intended primarily to complement the work with cascade-correlation presented in previous chapters. The original aim had thus been to re-use the experimental data used with cascade-correlation. However the sequence-learning tasks were considered unsuitable for use with skeletonisation for two reasons. Firstly, the ability of cascade-correlation networks to learn the counting task depends in part on the constructive aspect of that architecture. Secondly, the skeletonisation method was devised to act on non-recurrent backpropagation networks. Although recurrent architectures exist which are based directly on backpropagation (e.g., the SRN), it is unclear how the method for calculating relevance values would need to be adapted. In particular, in SRN's it is usually necessary to update weight-values after the presentation of each pattern, whereas relevance is calculated only every epoch. The experimental work in this chapter therefore focuses on the article-function experiment presented in chapter 5. Pilot studies showed that the full version of this task was difficult for the underlying backpropagation network to learn. The restricted version omitting the object-recognition component was thus used instead (see table 5.3).

7.4 Method

7.4.1 Training schedules and granularity

Two main approaches to scheduling incremental learning were examined. Mozer and Smolensky's original cyclic training and pruning schedule, which resembles cascade-correlation in the sense that a single unit is deleted at each phase, is compared with schedules in which the mechanisms used at each phase can be tailored in an attempt to give a closer fit to the profiles of the experimental data. These two sets of experiments also complemented each other in terms of the level of granularity at which phases of incremental learning could be said to correspond to phases in the RR model.

Mozer and Smolensky (1989b)'s training schedule consisted simply of a cycle through the following steps:

1. Train the underlying network using backpropagation, calculating relevance values at each unit

2. Delete the unit with the lowest relevance

3. Re-train the network on the same task using backpropagation

The number of times this cycle was repeated had to be pre-determined by the experimenter, although Mozer and Smolensky suggest that it would be possible to devise a stopping condition based on relevance values themselves (rather than simply an ordering on them).

7.4.2 Initial task training

Some difficulty was experienced in training the backpropagation network on even the restricted form of the task chosen. Variations in the parameters of backpropagation were first tried, but the network failed to converge even with learning rate and momentum set to very small values (e.g., 0.01) and even with a large maximum-epoch limit. Provision of large numbers of hidden units (e.g., 20 — recall that the data-set contained only 10 patterns) also failed to result in criterial performance.

These initial pilot studies showed that the network found particular patterns — those representing the two indefinite non-specific ambiguous cases — especially difficult to learn, and typically converged in as little as between 10 and 20 epochs when these were removed from the set. This observation prompted further pilot studies using training-set phasing to try to direct the network's learning towards correct performance on the original data-set. However, conservative incremental methods based on progressive increases in training-set size were also unsuccessful, regardless of the point at which the problematic patterns were introduced into the set. It was possible to obtain correct performance on these patterns if they were added relatively early, but only at the cost of performance on other classes of exemplar.

Finally an extended incremental training scheme was used, i.e., one in which training on a task which is not a subtask of the original task (as above) is used to stage learning of that task. It turned out that in this case such a task could be obtained by simply doubling the number of indefinite non-specific ambiguous exemplars from two to four to form a superset of the original data-set. The network was able to learn this twelve-pattern task in an average of 31.2 epochs. Retraining on the original ten-pattern dataset then rapidly converged (typically in only one further epoch). The effect of the extended set was simply that of increasing the statistical salience of the problematic class of cases. It is somewhat unfortunate that such engineering was necessary, as it changes the statistical profile of the task in ways not underlain by the empirical data in Karmiloff-Smith (1979a) (although frequency of each of these cases in everyday discourse is not discussed there — see note in section 5.2.1). In the studies which follow, preliminary training on the extended set preceded all phases of incremental training using skeletonisation.

7.4.3 Using Mozer and Smolensky's incremental training schedule

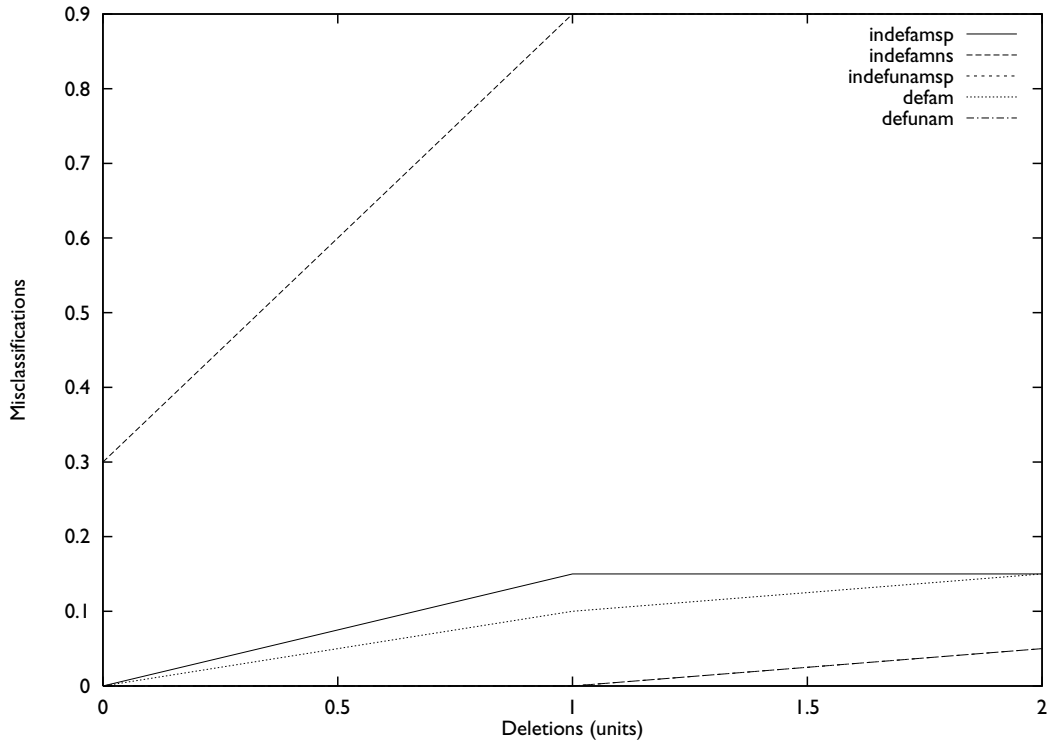
In these experiments the train–prune–retrain cycle was applied to both a network with three hidden units initially (which had been found during pilot studies to be the minimum number of hidden units needed to learn the task), as well as to a network with 6 hidden units. The second network was expected to form a redundant representation initially which would then become generalised through the action of skeletonisation as Mozer and Smolensky (1989b) describe.

Basic performance

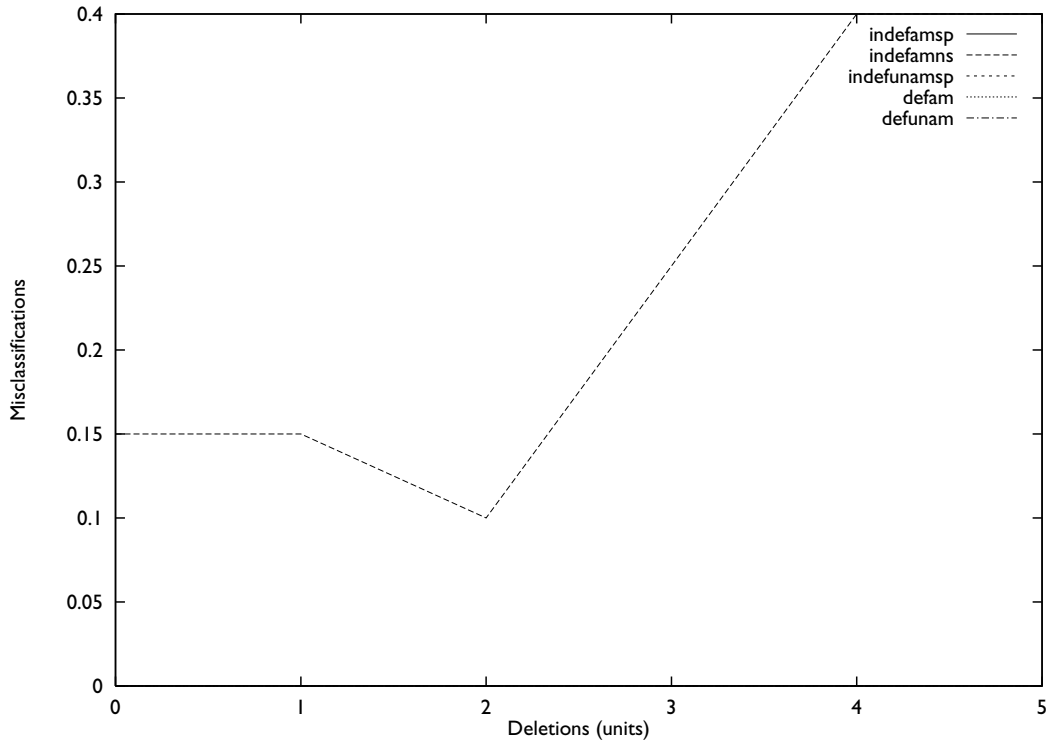
Training to mastery on the ten-pattern set required an average of 41.73 epochs (including preliminary training) in the networks initialised with three units, and an average of 19.55 epochs in networks initialised with six hidden units. The use of a larger number of hidden units than was needed to learn the task resulted in faster initial training as Mozer and Smolensky (1989b) note.

Skeletonisation according to the relevance measure After this initial training phase, the skeletonisation procedure was applied to networks with both sizes of hidden layer. In each case the unit with the lowest relevance at the end of each training phase was deleted. The stopping condition was that there must be at least one non-bias node remaining in order for it to remain a trainable three-layer network. The bias unit was also excluded from consideration for deletion. Figures 7.1(a) and 7.1(b) show how the proportion of misclassifications in each class of exemplar varied with successive rounds of relevance-based deletion.

As these figures show, the basic skeletonisation procedure does not result in U-shapes or fluctuations in the misclassification rate. Rather performance on particular classes of exemplars simply degrades as deletion progresses. However the pattern of misclassifications does resemble



(a) Network initialised with 3 hidden units



(b) Network initialised with 6 hidden units

Figure 7.1: Proportions of misclassifications on each class of exemplar for networks initialised with either (the minimum) 3 hidden units or 6 hidden units.

that in the empirical data in that misclassification rates on indefinite exemplars are consistently higher than those on the definite article.

Skeletonisation based on random selection of hidden units In the interests of investigating the effectiveness of the relevance measure as a means of selecting units for deletion, the above experiments were repeated using random selection of units to delete (again, excluding the bias unit and stopping with one non-bias unit remaining). Figure 7.2 shows how the proportions of misclassifications in each class varied with number of random deletions.

These figures indicate that the use of the relevance measure does have a significant guiding effect on incremental training beyond simply coalescing representations by reducing the dimensionality of the internal representational space, and acts to preserve correct performance whilst reducing network size as Mozer and Smolensky (1989b) intend. On the smaller network random deletion results in degradations in performance across the board (although the problematic indefinite non-specific ambiguous cases exhibit a peak in error reminiscent of the empirical data). The effect on the larger network appears more selective — performance on all three indefinite classes degrades (again monotonically) while performance on definite classes is spared.

To sum up, although this simple training schedule seemed to capture the decline in performance on the indefinite article it did not result in the subsequent improvement in performance associated with the U-shaped behavioural curve in this task. Mozer and Smolensky (1989b, p. 15) acknowledge that on some more complex tasks repeated skeletonisation may simply result in monotonically decreasing performance and it may be that the restricted form of the playroom experiment is an example of such a task.

7.4.4 Augmenting the basic skeletonisation scheme

In the light of the above results, the following studies were devoted to investigating whether augmenting the basic procedure with copying (as suggested by Clark and Karmiloff-Smith (1993)) and weight-freezing provided a better fit to the dynamics of the task.

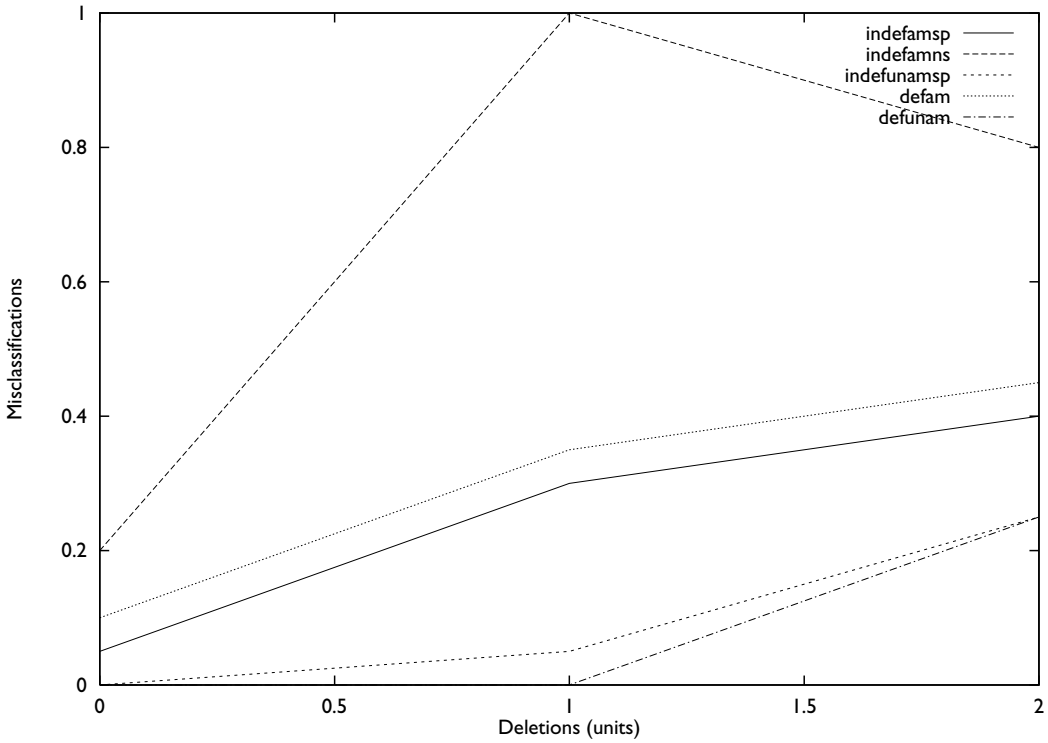
There are clearly a variety of possible copying strategies. For instance, it would be possible to use a vertical scheme in which extra layers were added after the original output layer or between the original hidden and output layers. In the experiments which follow, the decision was made to use a horizontal strategy in which the hidden layer was extended with extra units, with trainable connection to and from the output and input layers respectively. This arrangement allowed the network to select between forming a new task representation and making use of the previously learned one, rather than, as in the purely vertical case, having older structure bias all subsequent learning. The freezing procedure involved simply freezing the weights of a trained network. Freezing is motivated both by the desire to avoid catastrophic in subsequent training and also by the hypothesised conservatism of the redescriptive process.

Basic performance

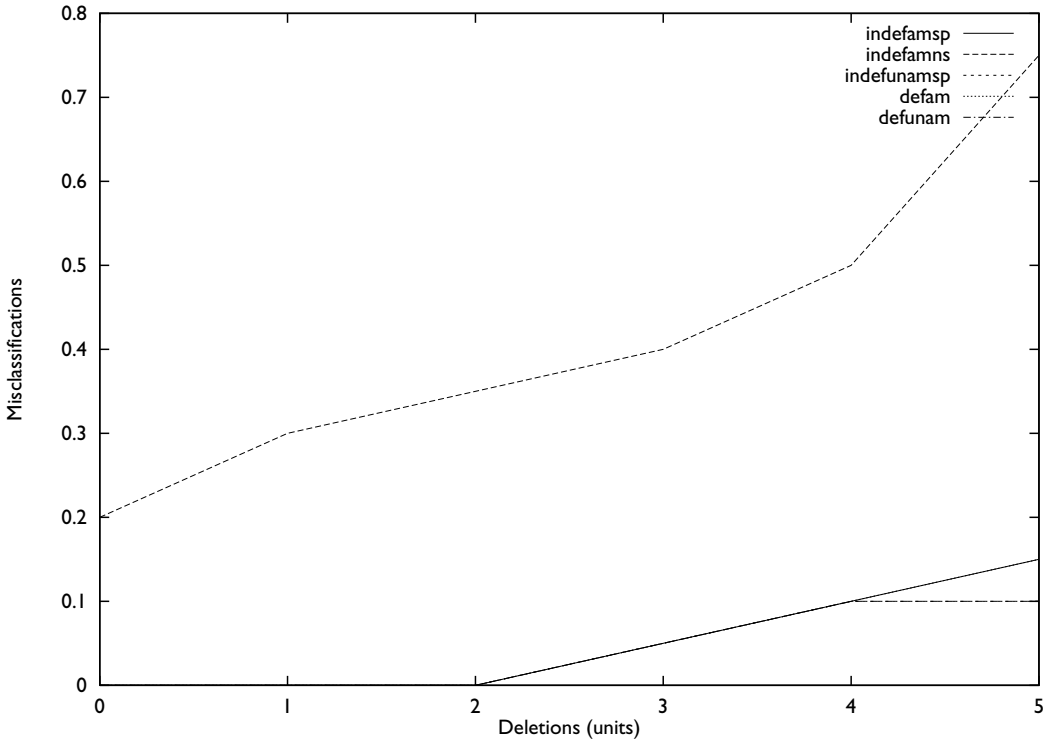
After the initial training phase, a network initialised with six hidden units was subjected to an augmentmented skeletonisation procedure. In this procedure, units were deleted according to relevance at the end of each training phase as before. The remainder of the network's trained weight structure was then frozen and either one, two or three new trainable hidden units added. The net was then retrained and the process repeated. For comparison, the process was repeated the same number of times as in the previous experiments, i.e., one less than the number of (non-bias) hidden units.

Figures 7.3(a), 7.3(b) and 7.3(c) show the misclassification rates by class for the three different rates of re-resourcing tried.

It had been hoped that by providing the network with additional trainable structure after deletion, that the drop in performance on the indefinite class could be reversed, giving a U-shaped behavioural curve as the data requires. However, as the figures show, re-resourcing the network did nothing to change the monotonic profile of the misclassification rates — as in the

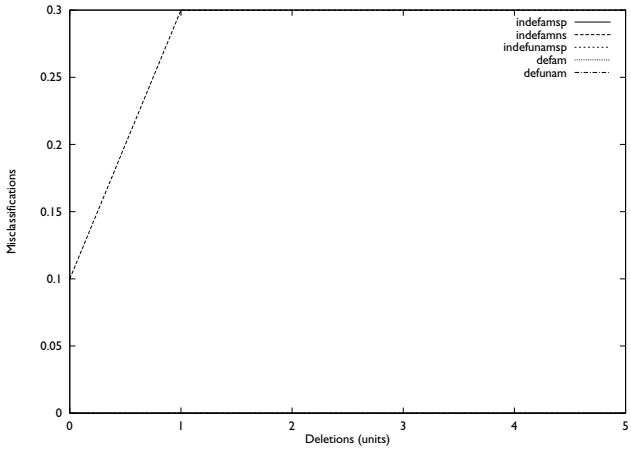


(a) Network initialised with 3 hidden units

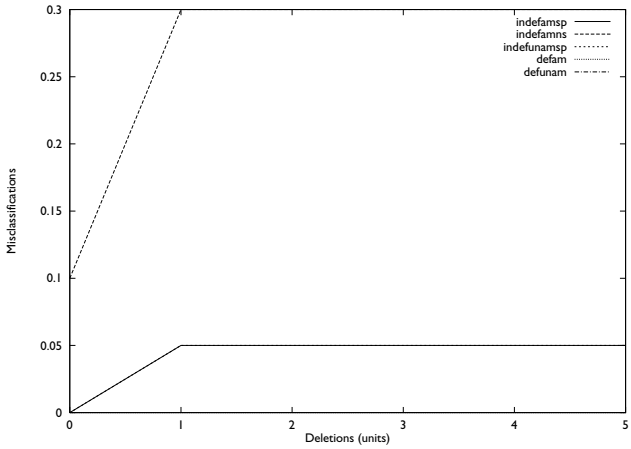


(b) Network initialised with 6 hidden units

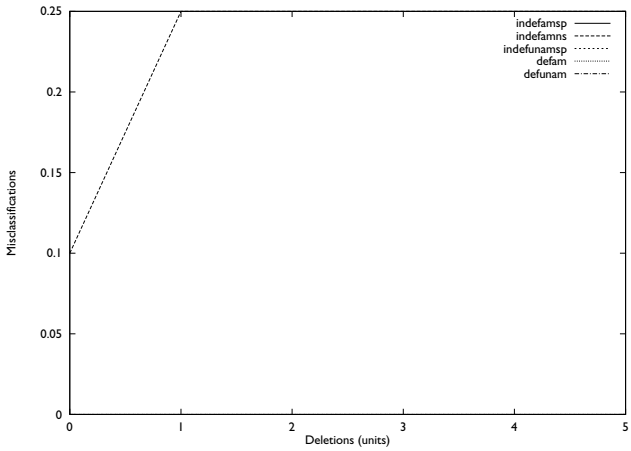
Figure 7.2: Proportions of misclassifications on each class of exemplar for networks initialised with either (the minimum) 3 hidden units or 6 hidden units and deleting hidden units randomly.



(a) Adding one extra hidden unit after each deletion

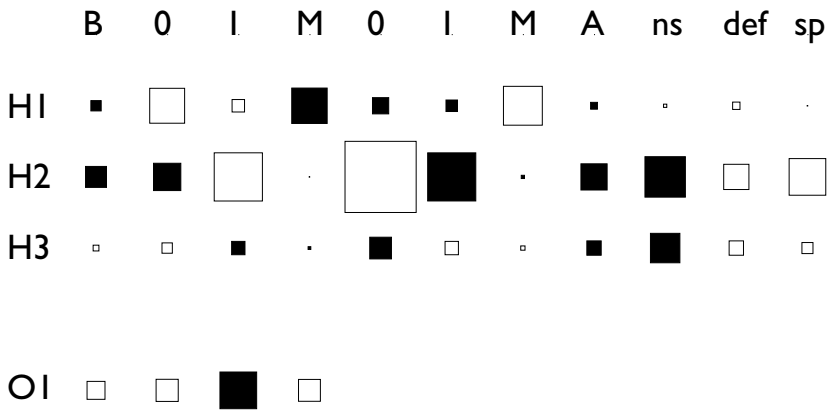


(b) Adding two extra hidden units after each deletion

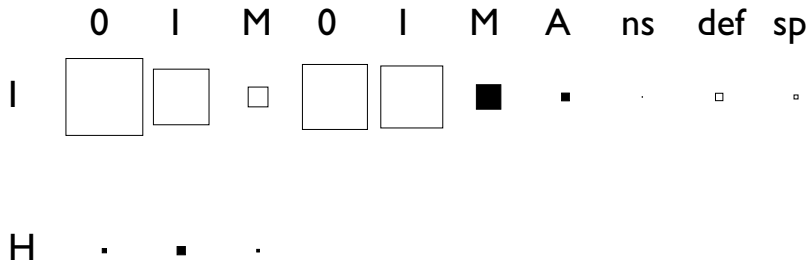


(c) Adding three extra hidden units after each deletion

Figure 7.3: Effect on misclassification rates of repeatedly adding different amounts of new trainable structure after network freezing



(a) Weights from a successful network



(b) Relevances from a successful network

Figure 7.4: Weight and relevance patterns for a backpropagation network trained successfully on the playroom task without object recognition

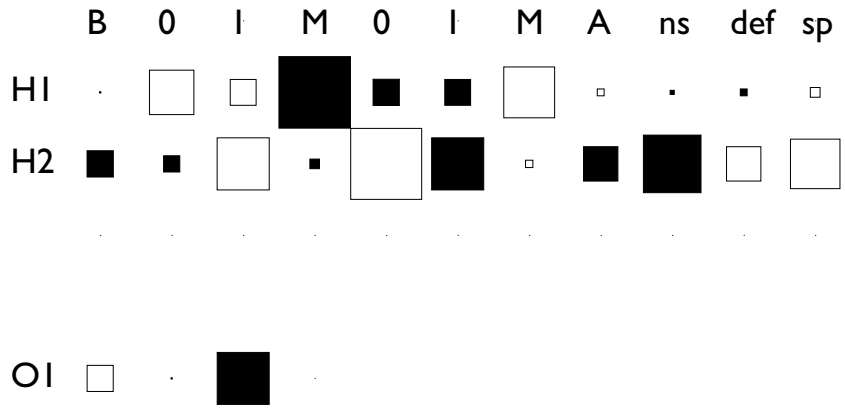
basic skeletonisation scheme, networks still either correctly classified exemplars from all classes throughout the training schedule, or else after deletion had given rise to misclassifications on particular classes, performance on this classes decreased monotonically as before.

Further pilot studies investigated the possibility that deletion could capture some of the effects of redescription after behaviourial mastery, while re-resourcing without further deletion might be better suited to promoting the subsequent improvement in performance at the transition to E2/3. However, no such schedule was found to exhibit patterns of performance other than the 'always-converge' and 'fail-after-initial-convergence' patterns seen in the studies above.

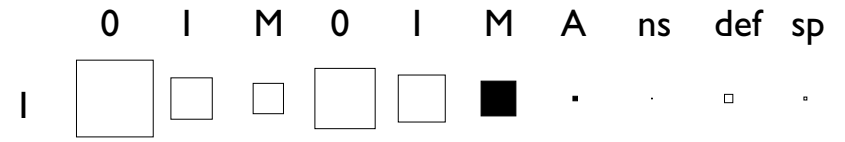
Analysis of representations

For the sake of comparison with the work on cascade-correlation, Hinton diagrams were plotted of the weights before and after deletion. Figures 7.4(a) and 7.4(b) show the patterns of weights and relevances (associated with units) for a three-hidden-unit network after convergence on the ten-pattern data-set, while figures 7.5(a) and 7.5(b) show the same network after hidden unit H3 has been deleted due to its having the lowest relevance.

These figures indicate that relevance-based deletion eliminates units responsible for correct performance on smaller, or less typical classes of exemplars, such as the indefinite non-specific ambiguous class.



(a) Weights after unit-deletion and re-training



(b) Relevances after unit-deletion and re-training

Figure 7.5: Weight and relevance patterns for a network which had started to misclassify indefinite non-specific exemplars after retraining following the deletion of the hidden unit with the lowest relevance value

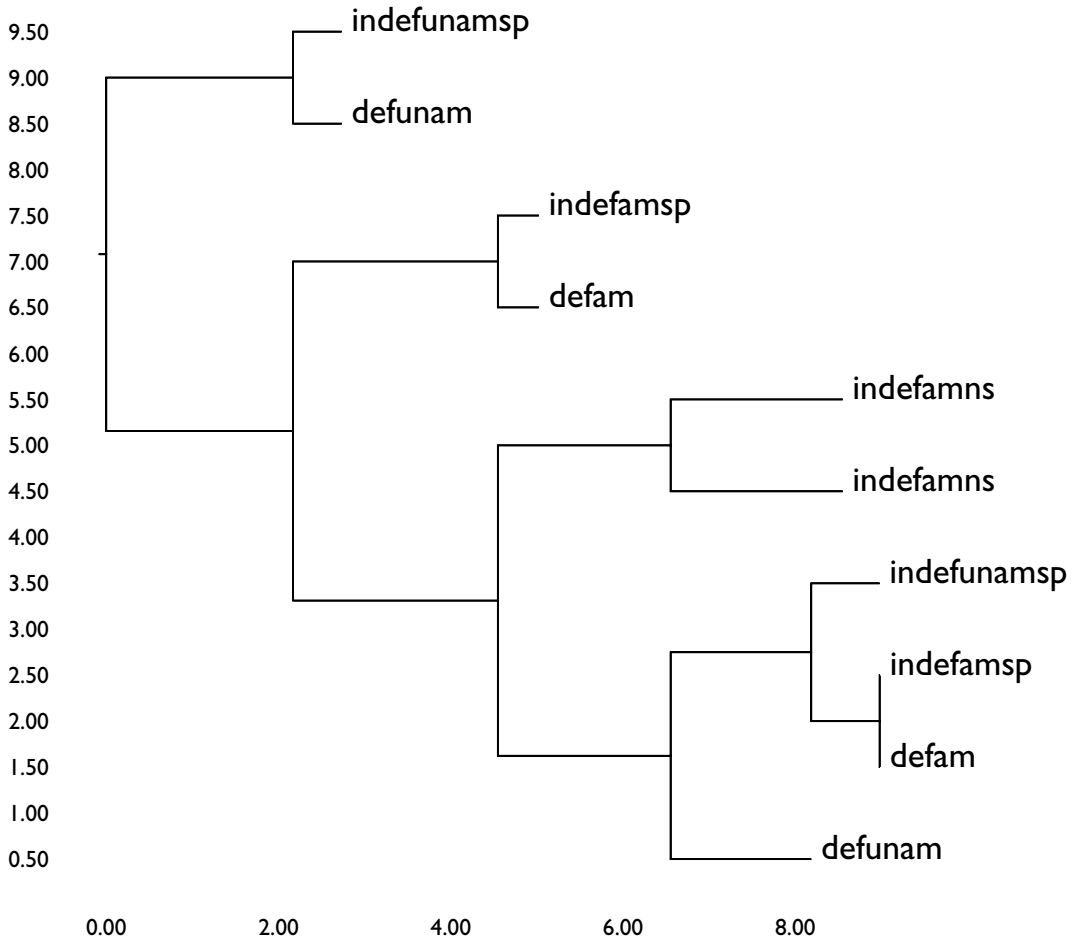


Figure 7.6: Cluster analysis of hidden-unit values after convergence on the ten-pattern data-set. Key: (un)am = (un)ambiguous, ns = non-specific, sp = specific, (in)def = (in)definite

7.5 Statistical analysis

Unlike cascade-correlation, since skeletonisation is performed on backpropagation networks which have a 'flat' hidden layer structure, it is possible to use statistical techniques such as principal components analysis (PCA) and hierarchical cluster analysis (see Everitt and Dunn (1991) for instance) to examine the internal representations formed.

7.5.1 Cluster analysis

Figures 7.6 and 7.7 show the results of applying cluster analysis to the values of the hidden units after the network had converged on the ten-pattern data-set and after deletion of the unit with the lowest relevance.

The groupings in figure 7.6 strongly suggest that (for eight of the ten examples) the task representation formed in the network does not correspond to the conception of the task as being classified primarily according to article and secondarily according to function. Rather there is a basic division between exemplars with ambiguous and unambiguous arrays (i.e., cases in which there is at least one object of a particular type in each playroom versus cases in which an object of that type appears only in one playroom respectively), although even this is violated by the two

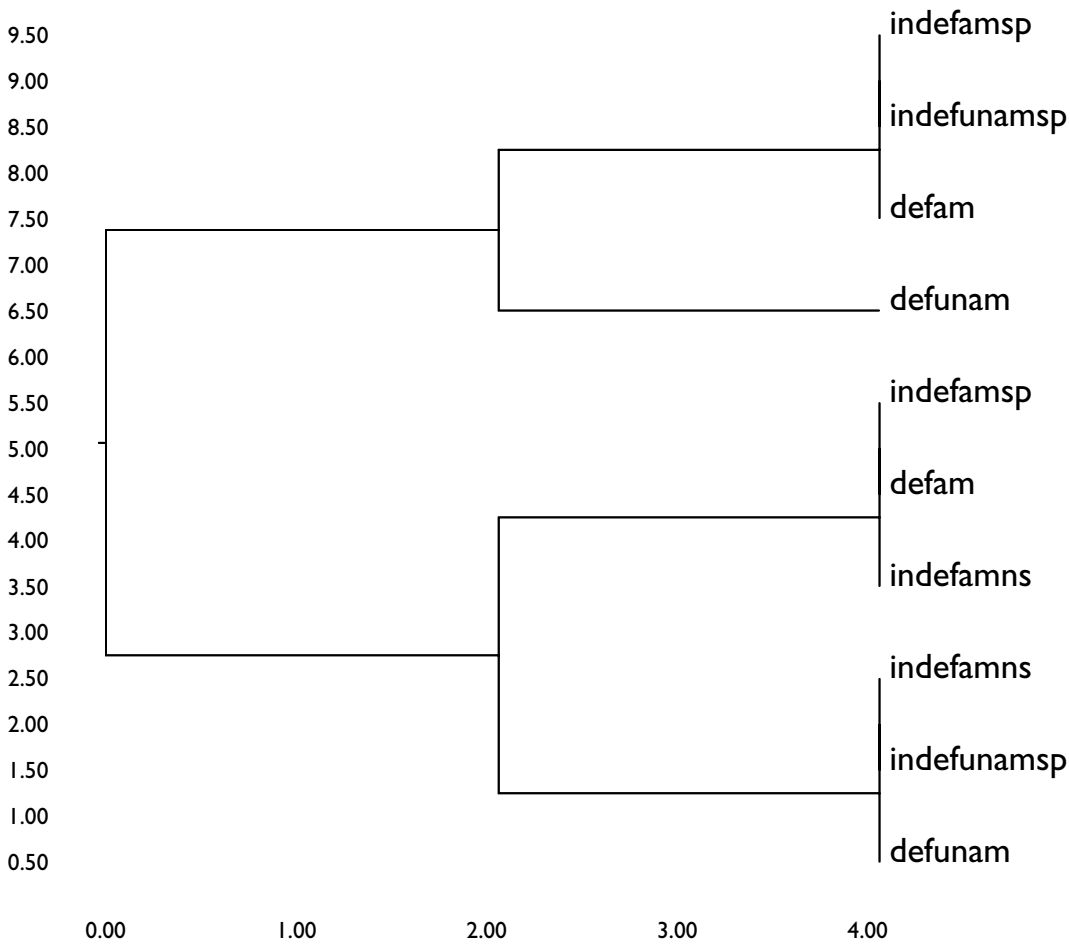


Figure 7.7: Cluster analysis of hidden-unit values after convergence on the ten-pattern data-set after deletion of unit with lowest relevance.

indefinite unambiguous specific exemplars. It seems that the groupings formed in the network's representation depend primarily on the quantity of an object and which playroom it appears in. It is interesting that the two indefinite non-specific ambiguous cases which necessitated the extended incremental training schedule constitute an exception to this pattern in that they appear close together in the diagram. The structure of figure 7.7 provides further evidence that the solutions found by error-driven learning centre around cues from the perceptual structure. This analysis shows that after the deletion of the unit supporting correct performance on the exceptional cases, the structure of the task representation is simplified and now depends in a simple manner on perceptual similarities.

7.6 Comparison with cascade-correlation

The difficulties encountered with training a backpropagation network to perform the playroom task without object recognition prompted a re-evaluation of the solution formed by cascade-correlation (see figure 5.11).

Examination of the pattern of weights formed by cascade-correlation suggested that performance on the first phase, in which all but the problematic indefinite non-specific ambiguous cases were correctly classified, was in fact underlain by a simple rule based on the pattern in the array alone (the weights from the article and function inputs were comparatively low). The role of the hidden unit was then to deal with the problematic cases, which constitute execeptions to the rule and for which article and function information must be attended to. Similar patterns of weights were to be observed in figure 7.4(a) suggesting that the backpropagation / skeletonisation scheme had developed a similar representation.

The pattern of repeated convergence followed by repeated failure without subsequent recovery over the course of skeletonisation is consistent with the effect of skeletonisation on the 'rule-plus-exception' example presented in Mozer and Smolensky (1989b, pp. 9–10). In this example a network with two hidden units is trained on 15 patterns which conform to a rule and a single exception. According to Mozer and Smolensky, the 'logical first candidate' for deletion is the hidden unit which has learned to treat the exceptional case. Although this behaviour is in keeping with the RRH in that it leads to greater generalisation with a possible loss in performance, and supports Mozer and Smolensky's claim that skeletonisation facilitates (experimenter) interpretation of network representations in terms of rules, it seems that the relevance measure here gives emphasis to essentially the same features as the statistical mechanisms of the underlying error-driven learning. Thus, in this case at least, the claims for relevance as a means of identifying non-statistical features of task structure seem somewhat weak.

7.7 Summary

This chapter has presented a small comparative study which examines the skeletonisation procedure applied to backpropagation networks as the basis for a model of representational redescription. Due to the comparative power of backpropagation and cascade-correlation and the lack of a version of skeletonisation adapted for recurrent networks the experiments focused on the form of the playroom experiment (chapter 5) omitting the object-recognition component.

Two main incremental training schedules were investigated. The first was simply to use the train–prune–re-train cycles used by Mozer and Smolensky (1989b). This was found to result in one of two behavioural profiles — networks either reconverged after every deletion without ever exhibiting a drop in performance, or else failed to converge after an initial run of successes and continued to fail thereafter.

For the second set of experiments, the basic skeletonisation scheme was augmented with two additional resource-phasing mechanisms — freezing of previously trained weight structure and addition of new trainable hidden units. The results of this second set of experiments were disappointing in that the addition of new structure did not facilitate a recovery in performance.

In conclusion, these studies suggest that unit deletion alone does not provide as good a fit to the experimental data as cascade-correlation, since although deletion was successful in causing the drop in performance on some indefinite-article exemplars, the network was never able to recover its performance on the exceptional indefinite non-specific ambiguous cases. Thus deletion alone seems unable to capture the redescriptive process at every phase of the RR model. As we saw above, the relevance measure may also not be as independent of statistical profiles as a model of the RRH would require in cases where the frequencies of significant exceptions are low.

Chapter 8

Discussion

This chapter summarises the experimental work presented in this thesis, discussing it in the context of the claims about connectionism which motivated it, as well as comparing it with related work. Possible directions for further work are than discussed as well as the general prospects for a connectionist model of representational redescription. Finally, conclusions are drawn from this project and its contribution summarised.

8.1 Summary of experimental work

8.1.1 Modelling plurifunctionality using cascade-correlation

The experiments presented in chapter 5 were designed to investigate whether cascade-correlation could be used to model the progression from individual procedural representations of the functions of the French article system to a systematic representation in which articles were represented plurifunctionally. The progression from unifunctional to plurifunctional representation is hypothesised to involve redescription of the representations of the article forms initially implicit in the separate unifunctional representations.

Karmiloff-Smith (1979a) tested children's comprehension of article-function using a setup in which, given a question, they had to use the article to determine to which of two dolls it was addressed when each doll was associated with a different configuration of objects. The simulation used a single (non-recurrent) cascade-correlation network, trained, as in the original experimental setup, to classify the utterances represented at the input layer according to which of the two dolls they would normally be addressed. In the test phase novel exemplars within the same setup were presented.

Network performance was measured using training-set error and generalisation using test-set error, both categorised according to the number of misclassifications on the particular combination of article, function and situation.

Results showed that although the network could learn the task, learning the association between the banks of input representations conveying object-types in the array and the object-information in the question had a significant effect on learning (although without this component the network converged after only one phase of input-side learning).

The aim that the network capture the overall behavioural profile (in particular the U-shaped curve on misclassifications of the indefinite article with the default non-specific function) was achieved, with an initial drop in errors across all categories on both training and test sets as the networks concentrated on learning the object-type mapping mentioned above, followed by some fluctuations in the proportion of misclassifications on different categories. This latter effect corresponded to the recruitment of individual (or small groups of) hidden units focused on correcting errors in each of the different categories. A simplified version of the experiment,

which omitted the object-recognition subtask, showed that the increase in representational capacity obtained through unit-recruitment was essential for correct performance on the indefinite article case.

Variation of internal parameters controlling the size of the search space (candidate pool size) and the duration of training in each phase (patience) was also investigated. It was found that training runs in which a large initial patience value was reduced according to the profile in figure 5.10(a) were most likely to exhibit an error profile resembling that of the original experiment, i.e., misclassification error on definite-article cases was consistently lower than that on indefinite cases, and the latter exhibited relatively large fluctuations in error (albeit never as great as those observed in children).

8.1.2 Modelling sequence learning using recurrent cascade-correlation

In this set of experiments, recurrent cascade-correlation was used to model redescriptive effects in sequence-learning domains. The RRH predicts several effects which apply across a range of such domains. In particular, redescription acts to individuate the components of sequences, and this effect begins with the ends of the sequence, progressing inwards as redescription takes place.

The experiments focused on a series of simple temporal counting tasks: counting with explicit markers, counting without explicit markers, which involved giving only the cardinality of the sequence as output, and a'comparative counting', task which required the network to count two successive sequences and then respond with 'true' if the first was longer than the second and vice-versa.

This set of tasks was used to investigate both the suitability of cascade-correlation for modelling redescription as well as the relationship between incremental learning and transferability of learning.

Although recurrent cascade-correlation networks were able to learn all three tasks, it was found that the extent to which sequence-boundaries were marked, and the timing of the required response, affected learning. The simulations captured several aspects of the RRH account of sequence-learning domains, in particular the fact that redescription proceeds from the ends of the sequence inwards and that the representations formed are initially sequential and become progressively less so. These progressions are seen both in the learning of a single network as well as over the course of training and transfer between series of tasks which require attention to increasingly non-sequential features.

Although transfer between the counting task and both the cardinal and comparative tasks was found to be positive, contrary to the predictions of the RRH, transfer from cardinality to the comparison task was negative.

As a complementary study to the counting experiments above, transfer between recurrent cascade-correlation networks was also investigated in the context of artificial grammar learning. Networks were trained to induce a finite-state machine and then required to transfer to another machine with identical structure. The aim of these experiments was to assess the extent to which cascade-correlation was bound by the perceptual structure of its input. Two transfer conditions were tried, corresponding to a relabelling of states and a relabelling of transitions. It was found that although transfer in the first condition did not require any new structure, in the second, transfer was uniformly negative, implying that the transfer task could make no use of the representations of the transitions embedded in the source.

8.1.3 Skeletonisation of backpropagation networks on article-function tasks

This complementary study investigated whether a selectionist resource-phasing scheme, such as the unit-pruning skeletonisation procedure could capture redescriptive effects. Two main incremental training schedules were tried, applied to the restricted form of the article-function task. The first was simply to use the train–prune–re-train cycles used by Mozer and Smolensky

(1989b), while the second augmented this with weight-freezing and the addition of new trainable hidden-unit structure.

The results of these experiments were particularly disappointing. Although the relevance measure was found to act selectively to preserve performance on the definite article cases, while producing a drop in performance on the indefinite article, it was found that, even using the augmented scheme, the network was not able to capture the subsequent increase in performance characterising the later part of the U-shaped behavioural curve in this micro-domain.

8.2 Cascade-correlation as a model of representational redescription

8.2.1 Cascade-correlation and the RR model

This section surveys the correspondence between cascade-correlation and the formats and phases of the RR model (as presented in section 2.2).

Innate constraints

Domain-general constraints As Karmiloff-Smith (1992a) argues, choice of connectionist architecture alone constitutes a basic kind of domain-general constraint. Thus the cascade architecture, and in particular its initial limitedness, are considered to act as domain-general constraints, as is the recurrent mechanism in the case of RCC.

Domain-specific constraints In the counting domain, the use of a discrete recurrent network was taken to be equivalent to the constraints of one-to-one correspondence, and item- and order-irrelevance. Parameter variation in the article-function experiments was also used to try to simulate the effects of early one-form–one-function constraints by controlling overfitting, with a degree of success. However in designing the input data format for the playroom experiment, a deliberate attempt was made not to bias the network towards forming a systematic representation of the articles and their functions.

The implicit level

As discussed in chapter 3, there is relative consensus among most commentators on the RRH that the implicit level of the RR model at least is relatively well captured by connectionist networks[1]. In forming distributed internal representations, cascade-correlation clearly exhibits the semantic opacity typical of connectionist models. However other aspects of level-I representations are not necessarily so naturally captured by cascade-correlation (or networks in general). For instance, one of the characteristics of level-I representations identified by Karmiloff-Smith is that at this level representations are added to the domain individually and without their commonalities being marked. Even though, as Plunkett and Marchman (1993) note, associative learning in conjunction with limited representational resources can lead to the rote-learning of a small number of representations, as the differences in misclassification by category in chapter 5 show. However the results of the parameter-variation studies suggest that it is difficult to enforce or control this kind of one-form–one-function constraint, as the experimental manipulations in the plurifunctionality study showed (chapter 5).

Counting tasks The initial phase of training in the counting domain resulted in a representation of number which was sequential and could not be interrupted, in that there was no way for a count to be started at any point but the beginning. This is consistent with the account of implicit-level representations of sequence-learning tasks.

Article-function task In the article-function task the implicit level was characterised by lack of systematicity (specifically 'unifunctionality'). Analysis of misclassifications provided evidence of such qualitative differences in the representation of items from different categories.

[1] Again, apart from the caveats made by Karmiloff-Smith (1992c, 1992a) regarding the tendency of connectionist models to ignore the innate constraints on the acquisition of such knowledge.

The E1 level

Article-function task Here level E1 was marked externally by a rise in misclassifications on the indefinite article. This effect was not reliably observed in the majority of the networks used, regardless of the proportions of definite and indefinite exemplars used. However, manipulation of candidate-pool size was found to produce such a profile with greater regularity.

Counting tasks Although, as noted in chapter 6, the counting tasks are not presented in terms of the three-phase RR model, the accessibility of representations formed during the counting with explicit markers task to the cardinality task would suggest that these representations are at level E1, since transfer is between tasks belonging to the same domain. Transfer between counting and cardinality was found to be positive in both time (number of epochs) and structure (number of hidden units). However the results on the complementary structured sequence-learning task suggested that this positive result relied upon the perceptual similarity of the tasks.

The E2/3 level

Many of the aspects of this level were not addressed by the simulations. In particular, the accessibility of representations to tasks in other domains was not investigated, and verbalisability was also considered to be outside the scope of this modelling effort. However, in the article-function tasks the E3 level is also characterised by a reconciliation of the conflicting article–function mappings which caused the rise in misclassification errors at level E1. At this general level, some cascade-correlation networks both reconciled the mappings (which was necessary for convergence) and captured the overall U-shaped behavioural profile.

Phase 1 (I–E1 transition)

One of the characteristics of this transition is a shift in the focus of attention from external inputs to internal representations. In the article-function experiments such a shift was to be observed in the relative strengths of weights from inputs to hidden units; initially the hidden units attended more strongly to the inputs, while subsequently recruited units attended more strongly to previously recruited hidden units. The drop in classification performance on the indefinite article tended to correspond to large shifts in the attention paid to different sources of information.

Phase 2 (E1–E2/3 transition)

On the article-function task the transition from level E1 to E3 is characterised behaviourally by a recovery in performance on the indefinite article. Corresponding to the shift in attention from input to internal representations, this transition is considered to involve renewed attention to external information in an attempt to reconcile this with the results of redescription during the previous phase. In the article-function experiments, although U-shaped curves are observed in some cases, no corresponding shift in attention is evident in the strengths of the weights into hidden units.

Although output-side training is designed to reconcile the structure newly added through input-side training, the results of the present studies did not seem to confirm the claims of Shultz (1994) regarding the correspondence between cascade-correlation and the RR model. The essence of his claim is that the initial error-driven learning phase corresponds to the initial learning to behavioural mastery in the RR model, the subsequent phase of correlation-driven learning to the shift in attention to internal information, and the eventual error-driven phase to the reconciliation of external and internal mappings in RR. Although aspects of this correspondence are borne out by some of the studies here, it is debatable whether in general single recruitment phases can be said to correspond to the RR model in this way. In almost all the cases considered here, several hidden units were recruited during each period which could be marked out as a qualitatively distinct behavioural phase. This finding is in keeping with those of Mareschal and Shultz (1993) who note that, in their cascade-correlation simulation of the balance scale, many of the unit-recruitments did not map directly onto more macroscopic strategy changes in network behaviour.

8.2.2 Roles of elements of cascade-correlation in modelling redescription

As we saw in chapter 4, there are general structural, procedural and behavioural similarities between cascade-correlation and both the RR process and model. The algorithm's hierarchical and conservative structure and its alternation of learning methods were the features given particular emphasis. This section surveys which of the features of cascade-correlation contribute most to its success at capturing RR. In the light of the experiments presented in chapters 5 and 6, the following conclusions can be drawn about the contribution of these aspects to cascade-correlation as a model of redescription, as well as other factors such as parameter manipulation.

Hierarchical structure

As expected, the hierarchical structure of the network architecture was found to give rise to effects on sequences similar to those required by the RRH (chapter 6). In particular, examination of weight-patterns showed that the features attended to by hidden units were initially sequential and became progressively less so, as more recently recruited hidden units attended to the lower-order results of previous learning. The ability to reuse the older feature-detectors upstream also manifested itself in the fact that an initial focus on the ends of sequences gave way to attention to groupings of interior elements.

Conservation of representations through weight-freezing

Clearly the preservation of previous learning through the freezing of input-side (input–hidden) weights also plays a role in producing the above effects. However, in section 2.6.2 doubts were raised concerning the domain-general status of such preservation of behaviours from previous stages — in particular it did not seem clear that it would be possible to elicit earlier behaviour in every domain associated with the RRH.

The freezing strategy of cascade-correlation also acted to give the fluctuations in misclassification error associated with the article-function mapping task in chapter 5. But as the studies of Squires and Shavlik (1991) and Mohraz and Protzel (1996) suggest, on some tasks freezing can be detrimental to both learning and generalisation performance, and it seems likely that freezing is partly responsible for the poor performance of the architecture on structural transfer tasks.

Learning mechanisms and granularity

Alternation of focus between error-driven and correlation-driven learning was found to act at too low a level of granularity to correspond to the macroscopic phase-progressions of the RR model. In all but the simplest cases (in particular the model of article–function mapping without object-recognition of section 5.4) several unit recruitments tended to correspond to a focus on a particular set of features or a trend in training or generalisation error. These findings run counter to the suggestion of Shultz (1994) that a single round of cascade-correlation learning (i.e., a phase of output-side learning, followed by a phase of correlation-driven learning and a second phase of error-driven learning) might correspond to the progression from I to E1 to E2/3 in the RR model. Rather, it seems that individual unit recruitments correspond to the view of the redescriptive process as something which acts recurrently at a microlevel.

Varying patience and candidate-pool size

The article–function mapping experiments of chapter 5 also investigated the effects of variations in two of the internal parameters of cascade-correlation — patience and candidate-pool size. In particular these parameters controlled the degree of overfitting of solutions and thus seemed to be useful in capturing both early constraints on the system such as the bias towards one-form–one-function mappings, as well as the tendency towards formation of separate and unsystematic representations of individual mappings characteristic of level I representations.

Correlation-driven learning

Although the correlation-driven (input-side) learning phase of cascade-correlation is neither directly error-driven nor directly within the network's input–output mapping, indirectly it is both

of these since it is driven by a function of the error which in turn relates to the input–output mapping being learned.

8.3 Exploring constraints on RR

8.3.1 Timing of redescription

Choosing to model redescription in the context of a supervised learning scheme, such as cascade-correlation or backpropagation, immediately raises the issue of whether redescription can occur before behavioural mastery, since in the terms of Clark and Karmiloff-Smith (1993) redescription is something assumed to succeed the characteristic error-driven learning of standard (first-order) networks.

Karmiloff-Smith (1994) is critical of what she sees as the assumption implicit in cascade-correlation that mastery is sufficient for redescription, implying that mastery is necessary but not sufficient. However, as we saw in Chapter 3, there is evidence that such reorganisations can occur before mastery (Goldin-Meadow & Alibali, 1994). Gentner et al. (1995) also focus on associative–relational change before mastery. Karmiloff-Smith (1992b) also sometimes talks of partial mastery or mastery of part of a task or domain. The idea of representational adjunction associated with level I makes it easier to imagine how parts of a particular task can be mastered piecemeal, their representations contributing incrementally to the stock which is eventually re-described and systematised.

To the extent that cascade-correlation can be considered to capture redescriptive effects, it suggests a picture of recurrent micro-redescriptions which accumulate to give the larger qualitative shifts in representation and behaviour associated with phase boundaries. It also suggests a redescriptive process which interpenetrates learning and is triggered by stability (success at subproblems) rather than requiring behavioural mastery.

8.3.2 Causes of redescription

There is some debate (e.g., Scutt and O'Hara (1993)) concerning whether representations formed in mastered tasks are spontaneously redescribed as Karmiloff-Smith claims or whether external factors such as pressure to improve overall performance or pressure to improve performance on new tasks are also involved. In cascade-correlation networks we know that any positive transfer must occur primarily without the influence of subsequent learning. This is because the input–hidden structure formed during the original training is frozen, and, as Pratt (1994) and Sharkey and Sharkey (1993) observe, these weights are significantly more important in transfer than the hidden–output weights, which are retrained with respect to the new mapping. (As noted elsewhere this also explains why transfer is poor in cases such as the grammar-learning experiments where the network cannot rely on changes in the hidden–output weights.)

8.3.3 Ordering of representational formats

The RR model proposes that knowledge passes through a series of representational formats in order of increasing explicitness and accessibility. The experiments in the number domain (chapter 6) investigated this ordering — transfer was carried out in both directions between tasks hypothesised to require degrees of explicitness. The results of these experiments were in agreement with the RRH that counting facilitated awareness of cardinality. However, transfer from cardinality to (ordinal) relations was uniformly negative, suggesting that the relations task did not rely on first forming an awareness of cardinality in this way, to the extent that it was unable to use the transferred representations (although it is possible that the incremental mechanisms in cascade-correlation give rise to a representation of cardinality which is then used in making comparative judgements).

8.4 Testing for RR

In general, it has been assumed here that what characterises redescribed knowledge is, at level E1, greater systematicity within the (micro-)domain, and, at higher levels, accessibility to processes associated with other domains. In practice the first of these was assessed via standard tests for generalisation, in particular performance on novel exemplars. Categorical misclassifications were also used as an index of systematicity in the article-function task. The assessment of the wider accessibility of representations to other tasks was restricted in practice by the domain-specific information available about transfer. Specifically, accessibility of knowledge of number could be investigated via transferability on a structural basis (or adaptive generalisation (Sharkey & Sharkey, 1993)) between the three tasks examined in chapter 6. However, for the article-function experiment, no data was available concerning the re-use of knowledge in subsequent learning on tasks even within the domain of language acquisition.

8.5 Cascade-correlation: conclusions

We are now in a position to assess how well cascade-correlation can be said to fulfill the specific requirements for a model of RR put forward in section 3.7:

- the model should treat its own representations as objects of manipulation
- do so independently of prompting by continued training inputs
- retain copies of the original networks
- form new structured representations of its own knowledge which can be manipulated, recombined and accessed by other computational processes.

As chapter 2 argues, the freezing strategy of cascade-correlation can be regarded as a way of preserving original learning. The question of whether it treats its own representations as objects of manipulation is more difficult to assess. Since connections are formed between each hidden layer and every previous hidden layer, subsequent learning is heavily mediated by previous learning, and it seems that it should also be possible for the network to select which of the frozen feature detectors to apply to the solution. However, as the results of the structured sequence learning experiments show, this selection mechanism is not powerful enough to extract structural information from a solution in such a way as to facilitate transfer to an isomorphic task. Thus, although it makes use of the results of previous training, it does not do so in such as way as to render them accessible and manipulable by other processes as the fourth requirement states.

8.6 Comparison of different schemes

8.6.1 Cascade-correlation and backpropagation

Cascade-correlation resembles backpropagation in the context of this discussion in being a multilayer, supervised learning scheme. The main differences between the two schemes are that backpropagation uses a single learning algorithm, is neither constructive nor preserves the results of previous training, and has both multi-unit hidden layer(s) and no cross connections.

The mechanisms giving rise to qualitative change in the two schemes are similar in being sensitive to training-set bias as Shultz et al. (1995) point out. Cascade-correlation's two-mode learning and constructive architecture does result directly in small qualitative changes, with error rising slightly after each unit recruitment as the error-driven output-side of the network accommodates to the new structure.

In the sequence-learning experiments the phasing of network resources led directly to the perceptual effects relating to sequence ends. In the plurifunctionality experiments the initial

limitedness of the network meant that performance again improved fastest on the most salient feature.

8.6.2 Cascade-correlation and skeletonisation

Although cascade-correlation is a constructive and skeletonisation a selectionist scheme, they cannot be regarded as directly complementary to each other. The main reasons for this relate to the differences discussed above between the cascade-correlation architecture and backpropagation. Other differences include the kind of off-line processing involved in each model. While cascade-correlation involves correlation-driven learning mediated by previous structure, skeletonisation works directly on the trained weights. In the terms of the discussion in Clark and Karmiloff-Smith (1993) and Bechtel (1993), cascade-correlation redescribes representations at the units, while skeletonisation acts on the procedure itself embodied by the weights, although there is some overlap in these procedures since skeletonisation deletes units rather than connections, and the new unit structure recruited by cascade-correlation is affected by the previously trained weights. Quartz and Sejnowski (forthcoming) also present recent evidence for the argument in favour of neural constructivism over selectionism as the predominant mechanism underlying representational change during cortical maturation.

8.6.3 Comparison with other work on explicitation

Greco and Cangelosi's redescription model

Although their model (see section 3.9.1) appears to capture the idea of a redescription process which acts entirely off-line to the usual error-driven input–output mapping, there are several aspects of the RR model omitted by Greco and Cangelosi (1996b) and which the present study addresses. Firstly, they assume that the explicitness of representations can be assessed through inspection of the results of unsupervised learning of categories from the hidden-layer representation of a backpropagation network. Accessibility of the resulting representations to other processes is not addressed in practice. Their work does not attempt to model tasks cited by Karmiloff-Smith, unlike the present study, and nor does it investigate the dynamics of change over a number of phases as this study does.

Similarities between this model and the cascade-correlation models include the freezing of the network structure embodying knowledge of the original task and the error-driven method used in the initial learning phase.

Thornton's explicitation model

Like the above model, this model incorporates non-error-driven learning, but, in its use of scaffolding through training-set change, inherently addresses knowledge reuse. The explicitation model differs from those presented here in that it is intended to investigate how incremental learning (as well as methods for detecting signs that a relational solution might be required) might be used to bring so-called 'hard problems' within the reach of statistical learners such as conventional neural networks. The present study is focused instead on investigating the use of incremental learning to capture redescriptive effects and overall dynamics in specific microdomains, rather than attempting to address whether or not these tasks could be learnt using a non-incremental scheme such as backpropagation.

8.7 Directions for further work

Although there are a number of specific extensions which could be made to each of the playroom and counting models, the suggestions below apply more generally to the modelling of the RRH using the incremental schemes covered here. The key motivations behind these suggestions continue to be improving the correspondence between the timing and granularity of qualitative change in the RRH and the network schemes, and the use of these schemes to explore possible

formal constraints on the RR process such as initial configurations and the relative influence of external and internal factors in causing change.

8.7.1 Variants on the cascade-correlation architecture

Extending the study of variation of internal parameters

Chapter 5 presented the results of an investigation into controlling qualitative behavioural and representational change through variation of the internal parameters of cascade-correlation, specifically patience and candidate-pool size, which affect the onset of changes between phases of differently directed kinds of learning and the extent to which representations are generalised respectively. The study of these parameters could be extended to consider their interaction, their effect in the context of other tasks and the effect they have on the transferability of representations.

It would also be possible to apply parameter-variation techniques to skeletonisation. For instance, in the skeletonisation scheme presented in chapter 7 the number of hidden units provided at each phase can be thought of as corresponding to the size of the candidate pool in cascade-correlation and could be manipulated in a similar way.

Using a more flexible unit-recruitment scheme

As we saw in chapters 5 and 6, qualitative shifts in behaviour often took place over the course of several unit-recruitments (see figure 5.6(b) for instance). It is possible that using a more flexible version of cascade-correlation such as Mohraz and Protzel (1996)'s FlexNet architecture, in which multiple recruitments may be made simultaneously into a single layer or into several new or existing layers, correlation-driven phases could be made to correspond more naturally to qualitative changes in the task to be modelled.

Combining aspects of cascade-correlation and skeletonisation

Although in the results presented in chapter 7 skeletonisation and copying of backpropagation networks appeared to be less powerful in general, there are aspects of that scheme which can be seen as closer to the spirit of the RRH than is cascade-correlation. For instance the skeletonisation scheme requires the network to reach behavioural mastery, and once this point is reached, acts off-line and using a method unrelated to the reduction of error. The pruned networks resulting from skeletonisation also capture the idea of reduced representations in a way in which cascade-correlation does not, except perhaps in the fact that the hidden–output mapping formed at earlier stages, and which adjusts performance to minor details, is subsequently lost.

It would be possible to incorporate certain aspects of the skeletonisation procedure in a model built in the cascade-correlation framework. In particular, pruning at each recruitment stage could be used to capture the idea of reduced representations. Skeletonisation itself does not seem to be directly applicable to the hidden-unit structure formed in cascade-correlation networks, since it involves deleting units, and cascade-correlation is already restricted to adding single-unit layers. However it would be possible to apply a pruning scheme to the layers of input–hidden connections after each round of input-side training. Wehrfritz (1994) presents a scheme based on cascade-correlation in which layers of input–hidden connections are pruned and which could form the basis of such a model.

Improving adaptive generalisation

It is clear from the grammar-learning experiments presented in chapter 6 that cascade-correlation performs poorly on transfer tasks when the perceptual aspects of the task are changed. In this specific case it would clearly be possible to address this by extending the network with an extra initial layer as in Dienes et al. (submitted). It is not clear that this approach would address the problems of adaptive generalisation in the general case.

An important extension to the current work would be the use of transfer to investigate accessibility of representations outside the original domain, i.e., in the terms of the RRH, level E2/3 representations.

The work of French (1995) and O'Reilly and McClelland (1994) has explored the use of twin-network schemes inspired by the hippocampus and neocortex to avoid catastrophic interference between sequentially learned concepts. It is possible that some of the techniques from these models could be incorporated in an improved model of transfer. However any further investigation of RR and transfer would also need to address the issue of whether transferable representations can be formed in a network trained on one domain and transferred to another domain without information from the second domain being used in any way to inform the design or training of the first network.

Extending the comparative study of error-driven models

Comparison with backpropagation It would be interesting to compare the performance of standard backpropagation with cascade-correlation on the training- and test-sets used here. This would substantiate the conjectures made above that the two algorithms capture similar qualitative-change phenomena via different means, i.e., via herding in backpropagation and freezing in cascade-correlation.

8.8 Contributions of this thesis

This thesis presents the first study dedicated to investigating the claims that connectionist architectures can provide models for the RRH in the context of particular domains discussed as evidence for RR effects by Karmiloff-Smith, specifically sequence-learning (exemplified by counting) and language acquisition. In particular it investigates whether a class of such architectures — those which are both incremental and error-driven — are particularly suited to this modelling effort. It is also the first practical investigation of network transfer as the operationalisation of the progressive accessibility characteristics of the RRH.

The playroom experiment extends the range of incremental learning techniques which have been used in developmental models based on cascade-correlation. Specifically, the patience and candidate pool-size parameters were varied over the course of training in an attempt to control the timing and nature of qualitative representational and behavioural change as well as to capture the early one-form–one-function constraint.

The study of counting, cardinality and comparisons was the first use of recurrent cascade-correlation in constructing a developmental model of temporal behaviour. The application of cascade-correlation to structural transfer between isomorphic but re-labelled finite-state machines was also novel.

The short study using skeletonisation was the first application of this technique in an attempt to model the RRH. The augmentation of the technique with weight freezing and network copying was a novel extension to skeletonisation.

8.9 Conclusions

There are several important reasons for concluding that cascade-correlation architecture as it stands does not constitute a model of representational redescription. In particular, experiments with task transfer have shown that although the networks are able to generalise adaptively to related tasks, the generalisations they develop cannot be considered to be structure transforming generalisations as Clark (1993a) requires. The algorithm's capability to model external indicators of redescription such as U-shaped behavioural curves was found to be similar to that of wholly error-driven methods such as backpropagation, despite being underlain by different mechanisms such as weight-freezing. The studies also confirm some of the doubts about primarily error-driven methods, even in conjunction with indirectly-driven methods such as maximising correlations.

Despite these negative general results, the modelling effort has also confirmed that the incremental scheme itself, as well as the variation of the patience and pool-size parameters, was useful in capturing domain-specific effects in the counting and article-function mapping tasks.

The results of the sequence-learning experiment also suggested that cascade-correlation was better placed to capture a more gradualistic RR model in which unit-recruitments correspond to micro-redescriptions and learning and redescription interpenetrate each other.

Bibliography

Abrahamsen, A. (1993). Cognizers' innards and connectionist nets: A holy alliance?. *Mind and Language*, *8*(4), 520–530.

Anderson, J. R. (1983). *The Architecture Of Cognition*. Harvard University Press, Cambridge, MA.

Basye, K., Dean, T., & Kaelbing, L. P. (1995). Learning dynamics: system identification for perceptually challenged agents. *Artificial Intelligence*, *72*, 139–171.

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

Bechtel, W. (1993). The path beyond first-order connectionism. *Mind and Language*, *8*(4), 531–539.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the Mind*. Blackwell, Oxford.

Bloom, P., & Wynn, K. (1994). The real problem with constructivism. *Behavioral and Brain Sciences*, *17*(4), 707–708.

Boden, M. A. (1990). *The creative mind : myths and mechanisms*. Weidenfeld and Nicholson, London.

Boden, M. A. (1988). *Computer Models of Mind: Computational Approaches in Theoretical Psychology*. C. U. P., Cambridge.

Boysen, S. T., & Berntson, G. G. (1995). Responses to quantity — perceptual versus cognitive mechanisms in chimpanzees (*Pan-Troglodytes*). *Journal of Experimental Psychology — Animal Behavior Processes*, *21*(1), 82–6.

Boysen, S. T., Berntson, G. G., Shreyer, T. A., & Hannan, M. B. (1995). Indicating acts during counting by a chimpanzee (*Pan-Troglodytes*). *Journal of Comparative Psychology*, *109*(1), 47–51.

Broadbent, H. A., Church, R. M., Meck, W. H., & Rakitin, B. C. (1993). Quantitative relationships between timing and counting. In Boysen, S. T., & Capaldi, E. J. (Eds.), *The Development of Numerical Competence: Human and Animal Models*, pp. 171–187. LEA, Hillsdale, NJ.

Brook, J. K. (1993). Developmental connectionism and representational redescription. In Brook, J. K., & Arvanitis, T. N. (Eds.), *The Sixth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex*, CSRP 300, pp. 26–31. School Of Cognitive and Computing Sciences, University Of Sussex, Brighton, UK.

Brook, J. K. (1995). Cascade correlation as a model of representational redescription. In Howell, A. J., & Wood, J. A. (Eds.), *The Eighth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex*, CSRP 390. School Of Cognitive and Computing Sciences, University Of Sussex, Brighton, UK.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Buckingham, D., & Shultz, T. R. (1994). A connectionist model of the development of velocity, time, and distance concepts. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 72–77 Hillsdale, NJ. Erlbaum.

Campbell, R. L. (1994). What's getting redescribed?. *Behavioral and Brain Sciences*, *17*(4), 710–711.

Carassa, A., & Tirassa, M. (1994). Representational redescription and cognitive architectures. *Behavioral and Brain Sciences*, *17*(4), 711–712.

Chrisley, R. L. (1993). Connectionism, cognitive maps and the development of objectivity. *Artificial Intelligence Review*, *7*, 328–354.

Chrisley, R. L., & Holland, A. (1994). Connectionist synthetic epistemology: Requirements for the development of objectivity. CSRP 353, School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton, UK.

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing*. MIT Press, Cambridge, MA.

Clark, A. (1993a). *Associative Engines: Connectionism, Concepts and Representational Change*. MIT Press, Cambridge, MA.

Clark, A. (1993b). Representational trajectories in connectionist learning. CSRP 292, School Of Cognitive and Computing Sciences, University Of Sussex, Falmer,.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language*, *8*(4), 487–519. (Originally published as CSRP 193, School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton, UK, 1991).

Clark, A., & Thornton, C. J. (1993). Trading spaces: Computation, representation and the limits of learning. CSRP 291, School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton, UK.

Clark, A., & Toribio, J. (1994). Doing without representing?. CSRP 310, School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton, UK.

Cleeremans, A. (1993). *Mechanisms of Implicit Learning*. MIT Press, Cambridge, MA.

Craven, M. W., & Shavlik, J. W. (1994). Using sampling and queries to extract rules from trained neural networks. In Cohen, W. W., & Hirsh, H. (Eds.), *Machine Learning: Proceedings of the Eleventh International Conference* San Francisco: CA. Morgan Kaufmann.

de Gelder, B. (1994). The risks of rationalising development. *Behavioral and Brain Sciences*, *17*(4), 713–714.

Dennett, D. (1993). Learning and labeling. *Mind and Language*, *8*(4), 540–548.

Dienes, Z., Altmann, G. T. M., & Gao, S.-J. (1995). Mapping across domains without feedback: A neural network model of transfer across domains. In Smith, L. S., & Hancock, P. J. B. (Eds.), *Neural Computation and Psychology: Proceedings of the 3rd Neural Computation and Psychology Workshop (NCPW3)* Berlin. Springer-Verlag.

Dienes, Z., Altmann, G. T. M., & Gao, S.-J. (submitted). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. . Submitted to *Cognitive Science*.

Donald, M. (1994). Representation: Ontogenesis and phylogenesis. *Behavioral and Brain Sciences*, *17*(4), 714–5.

Edelman, G. M. (1992). *Bright air, brilliant fire : on the matter of the mind*. Allen Lane, London.

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

Elman, J. L. (1990b). Representation and structure in connectionist models. In Altmann, G. T. M. (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, pp. 345–383. MIT Press, Cambridge, MA.

Elman, J. L. (1991). Distributed representations, simple recurrent networks and grammatical structure. *Machine Learning*, *7*, 195–225.

Everitt, B. S., & Dunn, G. (1991). *Applied Multivariate Data Analysis*. Edward Arnold, London.

Fahlman, S. E. (1988). An empirical study of learning speed in back-propagation networks. CMU-CS-88-162, School Of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Fahlman, S. E. (1991). The recurrent cascade-correlation architecture. CMU-CS-91-100, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. CMU-CS-90-100, School Of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Fodor, J. A. (1983). *The Modularity of Mind*. MIT Press, Cambridge, MA.

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

Fodor, J., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*, 183–204.

Forrest, S. (1991). *Parallelism and Programming in Classifier Systems*. Research notes in Artificial Intelligence. Pitman, London.

Frean, M. (1990). The upstart algorithm: A method for constructing and training feedforward neural networks. *Neural Computation*, *2*, 198–209.

French, R. M. (1995). Interactive tandem networks and the sequential learning problem. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* Cambridge, MA. MIT Press.

Fuson, K. C. (1988). *Children's Counting and Concepts of Number*. Springer Verlag, New York.

Fuson, K. C. (1992). Relationships between counting and cardinality from age 2 to 8. In Bideaud, J., Meljac, C., & Fischer, J.-P. (Eds.), *Pathways To Number: Children's Developing Numerical Abilities*. LEA, Hillsdale, NJ.

Gallant, S. I. (1993). *Neural Network Learning and Expert Systems*. MIT Press, Cambridge, MA.

Gelman, R. (1990). Structural constraints on cognitive development. *Cognitive Science*, *14*, 39.

Gelman, R., & Gallistel, C. R. (1978). *The Child's Understanding of Number*. Harvard University Press, Cambridge, MA.

Gentner, D., Rattermann, M. J., Markman, A., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In Simon, T. J., & Halford, G. S. (Eds.), *Developing Cognitive Competence: New Approaches to Process Modelling*, pp. 263–313. LEA, Hillsdale, NJ.

Goldin-Meadow, S., & Alibali, M. W. (1994). Do you have to be right to redescribe?. *Behavioral and Brain Sciences*, *17*(4), 718–719.

Greco, A., & Cangelosi, A. (1996a). Language, categorization, and representation: a pilot study using neural networks. DISA-96-1, Department of Anthropological Sciences, University of Genoa, Genoa, Italy.

Greco, A., & Cangelosi, A. (1996b). A representational redescription method using competitive learning. DISA-96-7, Department of Anthropological Sciences, University of Genoa, Genoa, Italy.

Grush, R. (1994). Beyond connectionist versus classical AI: A control theoretic perspective on development and cognitive science. *Behavioral and Brain Sciences*, *17*(4), 720.

Halford, G. S. (1993). *Children's Understanding: The Development of Mental Models*. Lawrence Erlbaum, Hillsdale, NJ.

Halford, G. S., Wilson, W. H., Guo, J., Gayler, R. W., Wiles, J., & Stewart, J. E. M. (1994). Connectionist implications for processing capacity limitations in analogies. In Holyoak, K. J., & Barnden, J. A. (Eds.), *Advances in Connectionist and Neural Computation Theory: Vol. 2. Analogical Connections*, pp. 363–415. Ablex Publishing Corporation, Norwood, NJ.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages and computation*. Addison-Wesley.

Inhelder, B., & Piaget, J. (1969). *The early growth of logic in the child*. Norton Library, New York.

Jackson, S. A., & Sharkey, N. E. (1995). Adaptive generalisation in dynamic neural networks. In Niklasson, L., & Bodén, M. (Eds.), *Current Trends in Connectionism*. Lawrence Erlbaum, Hillsdale, NJ.

Jacobs, R., Jordan, M., & Barto, A. (1991). Task decomposition through competition in a modular connectionist architecture: the what and where visual tasks. *Cognitive Science*, *15*, 219–250.

Johnson, M. H., & Karmiloff-Smith, A. (1992). Can neural selectionism be applied to cognitive development and its disorders?. *New Ideas In Psychology*, *10*(1), 35–46.

Jordan, M. I. (1986). Serial order: A parallel distributed processing approach. ICS Report 8604, University Of California, San Diego, CA.

Karmiloff-Smith, A. (1986). From metaprocess to conscious access: Evidence from children's metalinguistic and repair data. *Cognition*, *23*, 95–147.

Karmiloff-Smith, A., & Clark, A. (1993). What's special about the development of the human mind/brain?. *Mind and Language*, *8*(4), 569–581.

Karmiloff-Smith, A. (1979a). *A Functional Approach to Child Language: A study of Determiners and Reference*. Cambridge University Press, Cambridge.

Karmiloff-Smith, A. (1979b). Micro- and macro-developmental changes in language acquisition and other representational systems. *Cognitive Science*, *3*, 81–118.

Karmiloff-Smith, A. (1983). A new abstract code or the possibility of multiple codes?. *Behavioral and Brain Sciences*, *6*(1), 149–150.

Karmiloff-Smith, A. (1984). Children's problem solving. In Lamb, M. E., Brown, A. L., & Rogoff, B. (Eds.), *Advances in Developmental Psychology Vol. 3*. Lawrence Erlbaum, Hillsdale, NJ.

Karmiloff-Smith, A. (1988). The child is a theoretician, not an inductivist. *Mind and Language*, *3*(3), 183–195.

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, *34*, 57–83.

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

Karmiloff-Smith, A. (1992b). *Beyond Modularity: A developmental perspective on cognitive science*. MIT Press, Cambridge, MA.

Karmiloff-Smith, A. (1992c). Nature, nurture and *PDP*: *P*reposterous *D*evelopmental *P*ostulates?. *Connection Science*, *4*(3 & 4), 253–269.

Karmiloff-Smith, A. (1993). Self-organisation and cognitive change. In Johnson, M. H. (Ed.), *Brain Development and Cognition*, pp. 593–617. Blackwell, Oxford.

Karmiloff-Smith, A. (1994). Transforming a partially structured brain into a creative mind. *Behavioral and Brain Sciences*, *17*(4), 732–745.

Karmiloff-Smith, A., & Inhelder, B. (1975). If you want to get ahead, get a theory. *Cognition*, *3*(3), 195–212.

Karmiloff-Smith, A., & Johnson, M. H. (1994). Thinking on one's feet. *Nature*, *372*, 53–54.

Keil, F. C. (1990). Constraints on constraints: surveying the epigenetic landscape. *Cognitive Science*, *14*, 135–168.

Kirsh, D. (1991). Today the earwig, tomorrow man?. *Artifical Intelligence*, *47*, 161–184.

Klahr, D. (1995). Computational models of cognitive change: The state of the art. In Simon, T. J., & Halford, G. S. (Eds.), *Computational Models and Cognitive Change*. LEA, Hillsdale, NJ.

Kuscu, I. (1993). Developing a computational model of representational re-description. In Brook, J. K., & Arvanitis, T. N. (Eds.), *The Sixth White House Papers: Graduate Research in the Cognitive and Computing Sciences at Sussex*, CSRP 300, pp. 72–76. School Of Cognitive and Computing Sciences, University Of Sussex, Brighton, UK.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago.

Langley, P., Bradshaw, G., & Simon, H. (1983). Rediscovering chemistry with the BACON system. In Michalski, R., Carbonell, J., & Mitchell, T. (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Tioga, Palo Alto, CA.

Langley, P. (1987). A general theory of discrimination learning. In Klahr, D., Langley, P., & Neches, R. (Eds.), *Production System Models Of Learning And Development*, pp. 99–161. MIT Press, Cambridge, MA.

Lenat, D. B. (1982). AM: discovery in mathematics as heuristic search. In Davis, R., & Lenat, D. B. (Eds.), *Knowledge-Based Systems in Artificial Intelligence*. McGraw-Hill, New York.

Lenat, D. B. (1983). EURISKO: A program which learns new heuristics and domain conc epts. *Artificial Intelligence*, *21*.

Ling, C. X. (1994). Predicting irregular past tenses: Comparing symbolic and connectionist models against native english speakers. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 577–582 Hillsdale, NJ. Erlbaum.

Losonsky, M. (1994). Beyond methodological solipsism?. *Behavioral and Brain Sciences*, *17*(4), 723–4.

MacWhinney, B. (1991). *The* CHILDES *project: tools for analyzing talk*. Lawrence Erlbaum, Hillsdale, NJ.

Mandler, J. M. (1988). How to build a baby: On the development of an accessible representational system. *Cognitive Development*, *3*(2), 113–136.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, *99*(4), 587–604.

Mandler, J. M. (1993). On concepts. *Cognitive Development*, *8*, 141–148.

Mareschal, D., & Shultz, T. R. (1993). A connectionist model of the development of seriation. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* Hillsdale, NJ. Erlbaum.

McClelland, J. L. (1989). Parallel distributed processing: implications for cognition and devlopment. In Morris, R. G. M. (Ed.), *Parallel Distributed Processing: Implications for psychology and neurobiology*. Clarendon Press, Oxford.

McClelland, J. L. (1995). A connectionist perspective on knowledge and development. In Simon, T. J., & Halford, G. S. (Eds.), *Developing Cognitive Competence: New Approaches to Process Modelling*, pp. 157–204. LEA, Hillsdale, NJ.

Mézard, M., & Nadal, J.-P. (1989). Learning in feedforward networks: The tiling algorithm. *Journal of Physics A: Mathematical and General*, *22*(12), 2191–2203.

Mitchell, M., & Hofstadter, D. R. (1990). The emergence of understanding in a computer model of concepts and analogy-making. *Physica D*, *42*, 322–334.

Mohraz, K., & Protzel, P. (1996). Flexnet: A flexible neural network construction algorithm. In *Proceedings of the Fourth European Symposium on Artificial Neural Networks*.

Mounoud, P. (1982). Revolutionary periods in early development. In Bever, T. G. (Ed.), *Regressions in Mental Development: Basic phenomena and theories*. Lawrence Erlbaum, Hillsdale, NJ.

Mozer, M., & Smolensky, P. (1989a). Skeletonization: A technique for trimming the fat from a network via relevance assessment. In Touretzky, D. S. (Ed.), *Advances in Neural Information Processing Systems 1* San Mateo, CA. Morgan Kaufmann.

Mozer, M. C., & Smolensky, P. (1989b). Using relevance to reduce network size automatically. *Connection Science*, *1*, 3–16.

Mozer, M. C., & Bachrach, J. (1991). SLUG: A connectionist architecture for inferring the structure of finite-state environments. *Machine Learning*, *7*, 139–160.

Neches, R. (1987). Learning through incremental refinement of procedures. In Klahr, D., Langley, P., & Neches, R. (Eds.), *Production System Models Of Learning And Development*, pp. 163–219. MIT Press, Cambridge, MA.

Neches, R., Langley, P., & Klahr, D. (1987). Learning, development, and production systems. In Klahr, D., Langley, P., & Neches, R. (Eds.), *Production System Models Of Learning And Development*, pp. 1–53. MIT Press, Cambridge, MA.

Newell, A. (1988). *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.

Nolfi, S., Parisi, D., & Elman, J. L. (1994). Learning and evolution in neural networks. *Adaptive Behavior*, *3*(1), 5–28.

Olson, D. R. (1994). Where redescriptions come from. *Behavioral and Brain Sciences*, *17*(4), 725.

Omlin, C. W., & Giles, C. L. (1994). Extraction and insertion of symbolic information in recurrent neural networks. In Honavar, V., & Uhr, L. (Eds.), *Artificial Intelligence and Neural Networks: Steps toward Principled Integration*.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a tradeoff. PDP.CNS.94.4, Carnegie Mellon University.

Patterson, F. G., Patterson, C. H., & Brentari, D. K. (1987). Language in child, chimp and gorilla. *American Psychologist*, *42*, 270–273.

Peterson, D. (1993). The representational re-description hypothesis compared against two cases. CSRP 93-3, Cognitive Science Research Centre, University of Birmingham, Birmingham, UK. (Presented to the 1993 AAAI Spring Symposium on AI & Creativity).

Philips, S., Halford, G. S., & Wilson, W. H. (submitted). Representational redescription: From associative to relational systems..

Piaget, J. (1953). *The origin of intelligence in the child*. Routledge, London.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

Plunkett, K. (1993). Making nets work hard. *Mind and Language*, *8*(4), 549–558.

Plunkett, K., & Marchman, V. (1989). Pattern association in a back propagation network: Implications for language acquisition. CRL Technical Report 8902, Center for Research in Language, University Of California, San Diego, CA.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquistion. *Cognition*, *38*, 1–60.

Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. *British Journal Of Developmental Psychology*, *10*, 209–254.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, *48*, 21–69.

Pratt, L. Y. (1993). Discriminability-based transfer between neural networks. In Hanson, S. J., Giles, C. L., & Cowan, J. D. (Eds.), *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann, San Mateo, CA.

Pratt, L. Y. (1994). Experiments on the transfer of knowledge between neural networks. In Hanson, S. J., Drastal, G. A., & Rivest, R. L. (Eds.), *Computational Learning Theory and Natural Learning Systems: Volume 1: Constraints and Prospects*, pp. 523–560. MIT Press, Cambridge, MA.

Preece, P. F. W. (1980). A geometrical model of Piagetian conservation. *Psychological Reports*, *46*, 143–148.

Quartz, S. R., & Sejnowski, T. J. (forthcoming). The neural basis of cognitive development: A constructivist manifesto. *The Behavioral and Brain Sciences*.

Reed, R. (1993). Pruning algorithms — a survey. *IEEE Transactions on Neural Networks*, *4*(5), 740–747.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. In McClelland, J. A., Rumelhart, D. E., & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models*, pp. 216–271. MIT Press, Cambridge, MA.

Rumelhart, D. E., McClelland, J. A., & the PDP Research Group (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. MIT Press, Cambridge, MA.

Rutkowska, J. C. (1993). *The Computational Infant: Looking for Developmental Cognitive Science*. Harvester Wheatsheaf, Hemel Hempstead.

Rutkowska, J. C. (1994a). Can development be designed? what we may learn from the Cog project?. CSRP 370, School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton, UK.

Rutkowska, J. C. (1994b). Situating representational redescription in infants' pragmatic knowledge. *Behavioral and Brain Sciences*, *17*(4), 726–7.

Sanger, D. (1989). Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science*, *1*(2), 115–139.

Savage-Rumbaugh, E. S., Murphy, J., Sevcik, R. A., et al. (1993). *Language Comprehension in Ape and Child*. No. 3–4, Serial number 233 in Monographs of the Society for Research in Child Development.

Schmidt, W. C., & Ling, C. X. (forthcoming). A decision-tree model of balance-scale development. *Machine Learning*.

Scholnick, E. K. (1994). Redescribing development. *Behavioral and Brain Sciences*, *17*(4), 727–8.

Schyns, P. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, *15*, 461–508.

Scutt, T., & O'Hara, K. (1993). 3, 2, 1 . . . we have cognition. *Mind and Language*, *8*(4), 559–568.

Servan-Schreiber, E., Cleeremans, A., & McClelland, J. L. (1991). Graded state machines : The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, *7*, 161–193.

Sharkey, N. E., & Sharkey, A. J. C. (1993). Adaptive generalisation. *Artificial Intelligence Review*, *7*, 313–328.

Shultz, T. R. (1991). Simulating stages of human cognitive development with connectionist models. In Birnbaum, L., & Collins, G. (Eds.), *Machine Learning: Proceedings of the Eighth International Workshop* San Mateo, CA. Morgan Kaufman.

Shultz, T. R. (1994). The challenge of representational redescription. *Behavioral and Brain Sciences*, *17*(4), 728–729.

Shultz, T. R., Buckingham, D., & Oshima-Takane, Y. (1994). A connectionist model of the learning of personal pronouns in English. In Hanson, S. J., Petsche, T., Kearns, M., & Rivest, R. L. (Eds.), *Computational learning theory and natural systems, Vol 2: Intersection between theory and experiment*, pp. 347–362 Cambridge, MA. MIT Press.

Shultz, T. R., & Elman, J. L. (1994). Analyzing cross connected networks. In Cowan, J. D., Tesauro, G., & Alspector, J. (Eds.), *Advances in Neural Information Processing Systems 6*, pp. 1117–1124 San Francisco, CA. Morgan Kaufmann.

Shultz, T. R., & Schmidt, W. C. (1991). A cascade-correlation model of balance scale phenomena. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pp. 635–640 Hillsdale, NJ. Erlbaum.

Shultz, T. R., & Oshima-Takane, Y. (1994). Analysis of unscaled contributions in cross connected networks. In *Proceedings of the World Congress on Neural Networks, Vol. 3*, pp. 690–695 Hillsdale, NJ. Lawrence Erlbaum.

Shultz, T. R., Schmidt, W. C., Buckingham, D., & Mareschal, D. (1995). Modelling cognitive development with a generative connectionist algorithm. In Simon, T. J., & Halford, G. S. (Eds.), *Developing Cognitive Competence: New Approaches to Process Modelling*. LEA, Hillsdale, NJ.

Simon, T. J., & Halford, G. S. (1995). Computational models and cognitive change. In Simon, T. J., & Halford, G. S. (Eds.), *Developing Cognitive Competence: New Approaches to Process Modelling*, pp. 1–29. LEA, Hillsdale, NJ.

Smith, L. B., & Jones, S. B. (1993). Cognition without concepts. *Cognitive Development*, *8*, 181–188.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*, 159–216.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*, 29–56.

Squires, C. S., & Shavlik, J. W. (1991). Experimental analysis of aspects of the cascade-correlation learning architecture. Machine Learning Research Group Working Paper 91-1, University of Wisconsin-Madison, WI.

Strauss, S., & Stavy, R. (Eds.). (1982). *U-Shaped Behavioral Growth*. Academic Press, London.

Thelen, E., & Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA.

Thornton, C. J. (1994). Unsupervised learning with the soft-means algorithm. In *Proceedings of the World Congress on Neural Networks. Vol. 4*, pp. 200–205.

Thornton, C. J. (1995). Brave mobots use representation. CSRP 401, School Of Cognitive and Computing Sciences, University Of Sussex, Brighton, UK.

Thrun, S., & O'Sullivan, J. (1995). Clustering learning tasks and the selective cross-task transfer of knowledge. CMU-CS-95-209, School of Computer Science, Carnegie Mellon University, Piitsburgh, PA.

van der Maas, H. L. J., & Molenaar, P. C. M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review*, *99*(3), 395–417.

van Gelder, T. (1992). What might cognition be if not computation?. Research Report 75, Cognitive Science, Indiana University.

Vinter, A., & Perruchet, P. (1994). Is there an implicit level of representation?. *Behavioral and Brain Sciences*, *17*(4), 730–1.

Wallace, I., Klahr, D., & Bluff, K. (1987). A self-modifying production system model for cognitive development. In Klahr, D., Langley, P., & Neches, R. (Eds.), *Production System Models Of Learning And Development*, pp. 359–436. MIT Press, Cambridge, MA.

Wehrfritz, C. (1994). Neuronale Netze als statistische Methode zur Erklärung von Klassifikationen. Master's thesis, Friedrich-Alexander-Universität, Erlangen-Nürnberg.

Wickelgren, W. A. (1993). Chunking, familarity and serial ordering. In Boysen, S. T., & Capaldi, E. J. (Eds.), *The Development of Numerical Competence: Human and Animal Models*, pp. 245–268. LEA, Hillsdale, NJ.

Wiles, J., & Bloesch, A. (1992). Operators and curried functions: training and analysis of simple recurrent networks. In Moody, J. E., Hanson, S. J., & Lippmann, R. P. (Eds.), *Advances in Neural Information Processing Systems 4*, pp. 325–332 San Mateo, CA. Morgan Kaufmann.

Wiles, J., & Elman, J. L. (1995). Learning to count without a counter: A case study of dynamics and activation landscapes in recurrent networks. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* Cambridge, MA. MIT Press.

Yamauchi, B., & Beer, R. (1994). Integrating reactive, sequential, and learning behavior using dynamical neural networks. In *Third International Conference on Simulation of Adaptive Behaviour*.

Young, R. M. (1976). *Seriation by Children: an Artificial Intelligence Analysis of a Piagetian Task*. Birkhauser, Basel.