Cascade Correlation as a Model of Representational Redescription

[illegible] Brook

CSRP [illegible]

July 199[illegible]

ISSN [illegible]

UNIVERSITY OF

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, oPn

Cascade Corre at on as a Mode of   epresentat ona edescr pt on

J K Broo

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 redescription 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

Contents

Cascade correlation and

Cascade correlation and plural functionality

1 . Exa p e do a ns

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 ev-

C apter —2

e epresentat ona edescr pt on Hypot es s

—2 Introduct on t e epresentat ona edescr pt ona Hypot es s

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.

—2ı I p c t and exp c t representat ons

Representations, in the terms of the RRH, are considered to be that which sustains behaviour

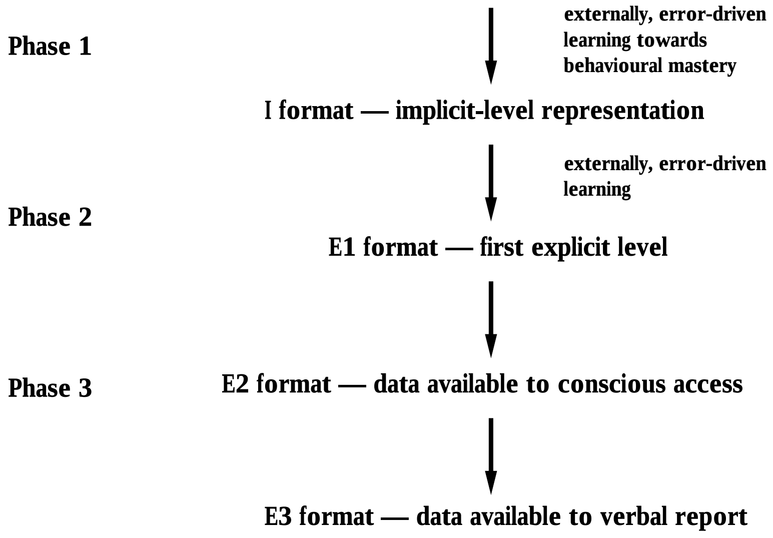

Phase 1
externally, error-driven learning towards behavioural mastery
I format — implicit-level representation
externally, error-driven learning
Phase 2
E1 format — first explicit level
Phase 3
E2 format — data available to conscious access
E3 format — data available to verbal report

Figure 2.1: The Representational Redescription Model

1993). The hypothesis also has it that in this phase new knowl

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 e H n 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.

21 e os t on of t e H

The RRH is offered as a speculative theory and a more or less implicit challenge to other theo-

Karmiloff-Smith (1994, p. 738) responds that the RRH has never denied that literacy training during development affects brain configuration. She disput

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

behavioural components in a supporting environment.

Other issues also remain to be addressed. For instance what p

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

whmw 1(s)o.4.281(s)3.19167(g)-3.19mglof5246.861(6(f)f.97311)-23f.97311lo,

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

—2—2Conservat on of ear er representat ons 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, Extent of redescr pt on

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 Ot er responses to t e H

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.

—2ı For of t e ode

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]

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}—^{2}Nature of representat ona for ats

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

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 be-

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 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.

C apter

Connect on s and Deve op enta Mode n

1 Introduct on

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.

—2 Co putat ona ode s of deve op ent

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

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.

—22Dyna ca Mode s

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.

—2 Connect on st ode s

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.

—2 D scuss on

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

> 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)

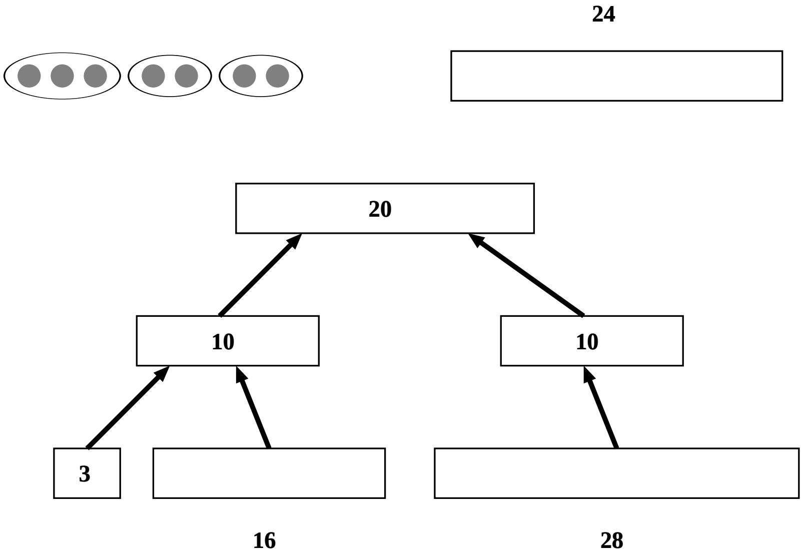
and concluding that

> 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

- **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 wh**


24
20
10
10
3
16
28

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

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 algorith

Chapter

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 learnimed bmemta e-re-correlation asaB(e)10.561y07(h)4.
fi.972(s)2.39793(t)-5.76685((k)4.973119474(i97384(e)10.56584(o)4.97311(s)2.396u1(e)32.4718(v)-5.674423(e)10.5616(-)-5.765(c)10.561
8641(e)10.5616(n)4.973u6(v)40.6433(e)10638(t)-5.7678.041(i)7.9853(n)4.973474(s)2.39682]TJT*[(311(f)-224.955(t)-2032((e)10.5616(d
8Td[(E)13.h es.6297(o)4.95845(e)10.5605(t)-5.76g irencrAs aFoo3(o)4.47311(a)-3.19279(r)46.2.61(d)4.97311(e)32.4718(310.569682(t)-

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.

e pro se of cascade corre at on as a ode of

The relevance of using cascade-correlation in the construc

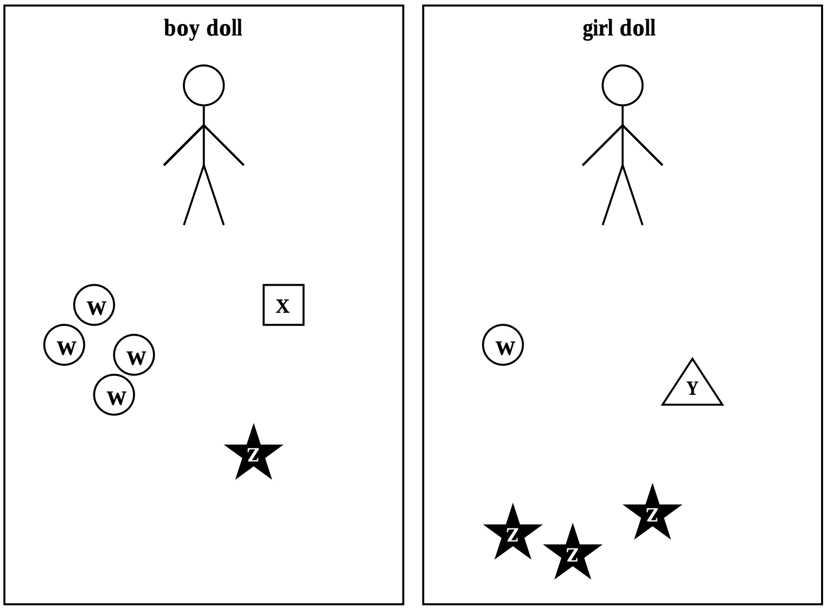


Figure 5.1: The experimental array used in the playroom experiment (after Kar

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

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

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.

.!

—2 est data

In order to test the generalisation of the learning of the different semantics for the indefinite article over the course aaaa

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.

, esu ts

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.

, 1 Bas c perfor ance

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.

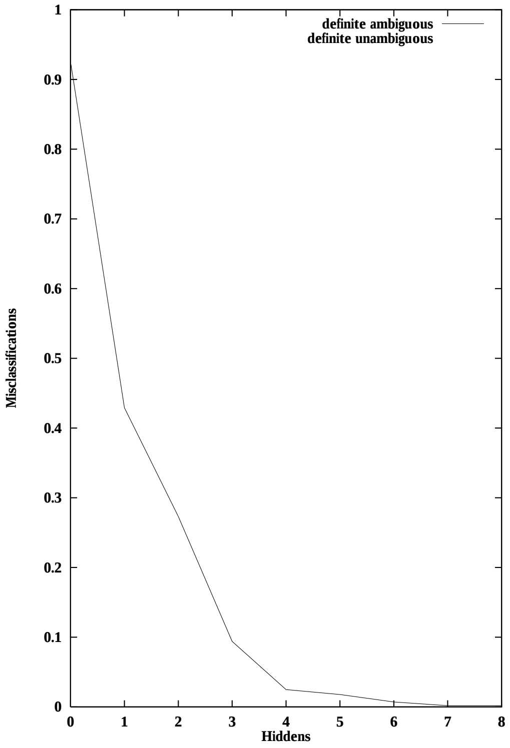
, —^{2}M sc ass cat ons

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.

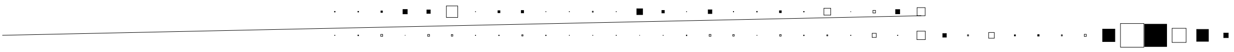
Misclassifications on training set Figure 5.4 shows the misclassifications across the difff25cg s. din

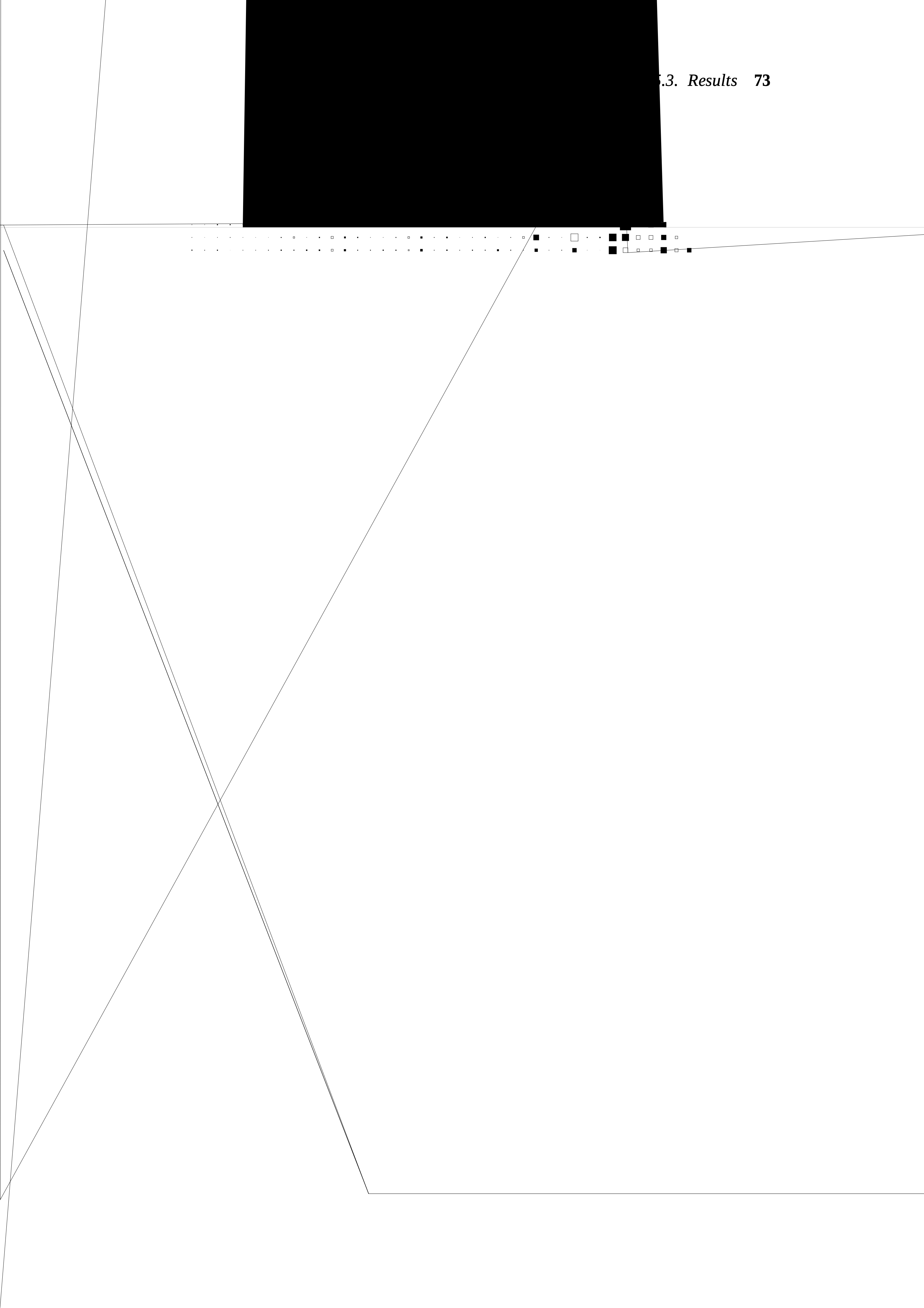
Definite article	**50%**		**ambiguous**	**25%**
			unambiguous	**25%**
Indefinite article	**50%**	**non-specific**	**ambiguous**	**25%**
		specific	**unambiguous**	**12.5%**
			ambiguous	**12.5%**

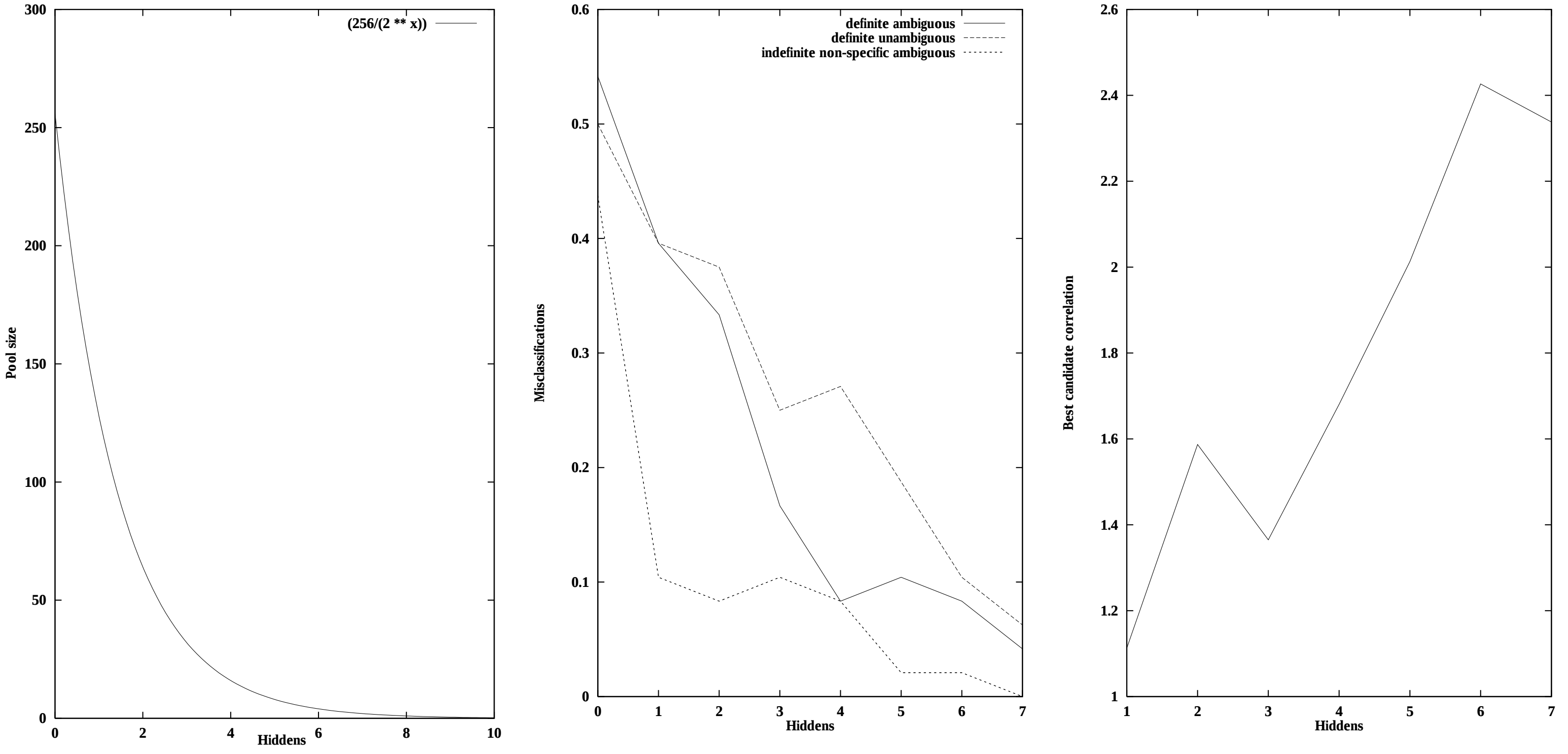
Input Epochs	Output Epochs	Average Hiddens	Average Epochs	Min/Max
150	15			


definite ambiguous
definite unambiguous
Misclassifications
Hiddens
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
6
7
8

0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6







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

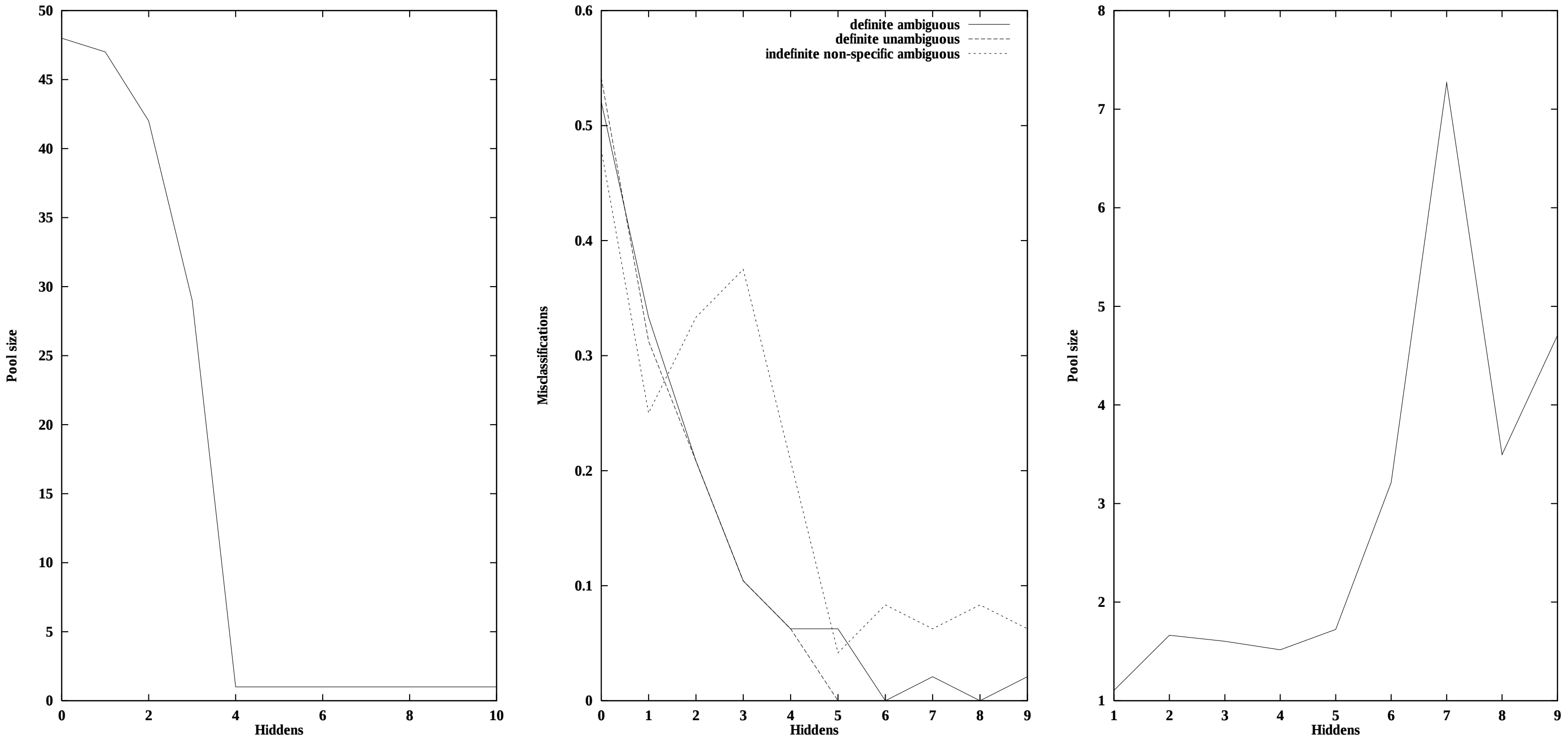
(b) Test-set misclassifications

(c) Correlation-measure *S*

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

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

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**



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

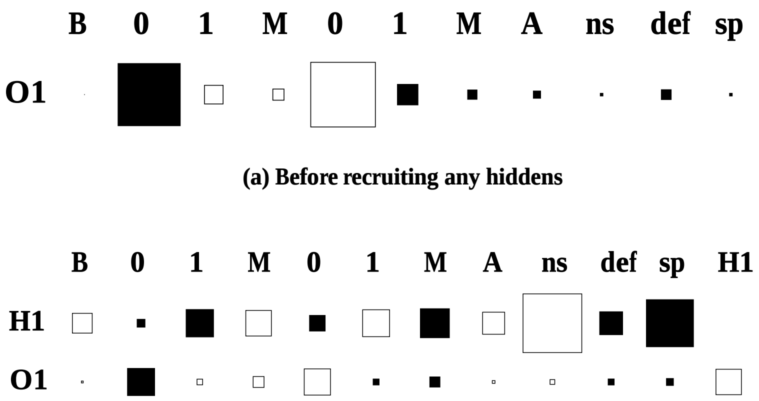
(b) Test-set misclassifications using (configuration C)

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

Figure 5.10: Effects of using a function of hidden-unit numb

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

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

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 exp

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 strengt

Chapter

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

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 usin

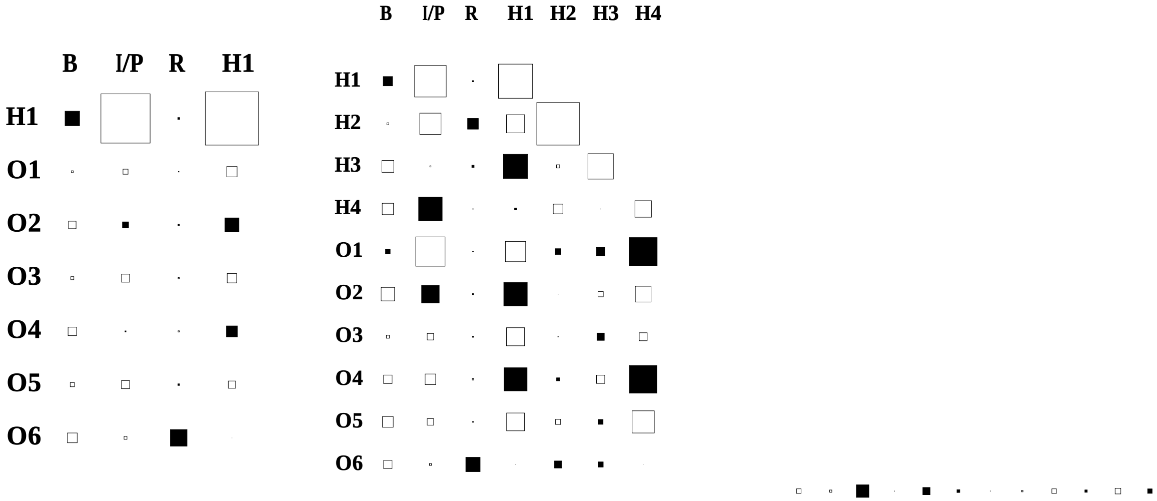
	A		**B**		**C**	
time	**input**	**output**	**input**	**output**	**input**	**output**
t_0	○	**1**	○	**1**	○	**1**
t_1	●	**2**	●			

800
700
600
500
400
300
200
100
0

Input Epochs	**Output Epochs**	**Average Hiddens**	**Average Epochs**
200	**200**	**3.7 (3/5)**	

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 (cardinalitE14(x111(t)1001(L)3.98(n)-2e)(o-.87n)11(2511(n)11(g6



(a) 1 hidden unit

(b) 4 hidden units

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)

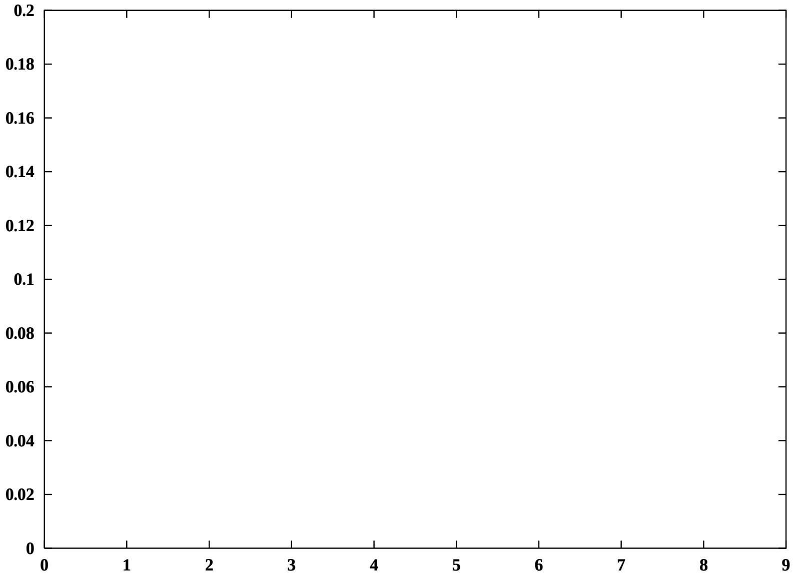
(b) Training set of 16 relations

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

3(g19.5357(t)1(r.7663)(c)1f.5616(5th)4.9b4(m)4.97237e r294316(r)2p9.5357(t)1(d7669861t(1v9.53ff7(t)-5t7F67(t)-95-qcfin375.803(d-244.29


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
1
2
3
4
5
6
7
8
9

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 exp

i72(t)-5.76573(s)2419278(t)-538.685(l36.123(i)7.98641(n)4.97423(p)4.9.39682(i)a76796(e1i)7.98641(n)4.97311(2(g)-354

ex

Epochs from	**Hidden units**	**Epochs after**	**Hidden units**		
random start	**from random**	**transfer**	**after transfer**	τ	

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

D scuss on

ı Co par son between perfor ance of CC and t e 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 o[illegible]

were not restricted to being loc wTait9(W)59.383(i)7.98641(l)7.98641(e)10.56796(s)4.986

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

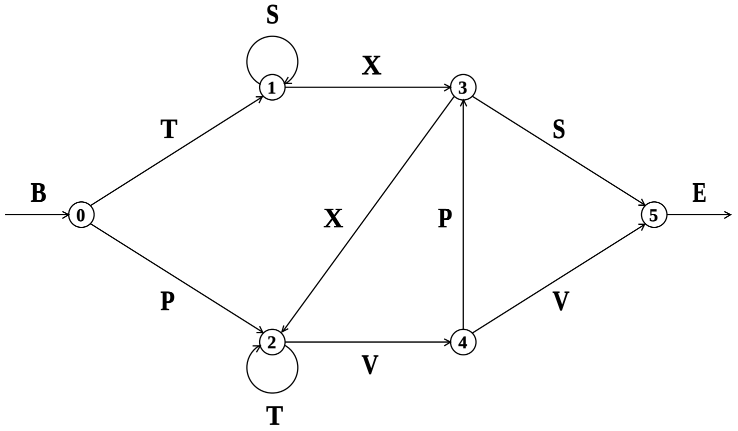


Figure 6.11: The finite-state machine accepting the Reber grammar

ı tructura transfer between so orp c ac nes

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(a)-3.18832(r)98641(r)24.33.18dr

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 ag

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.

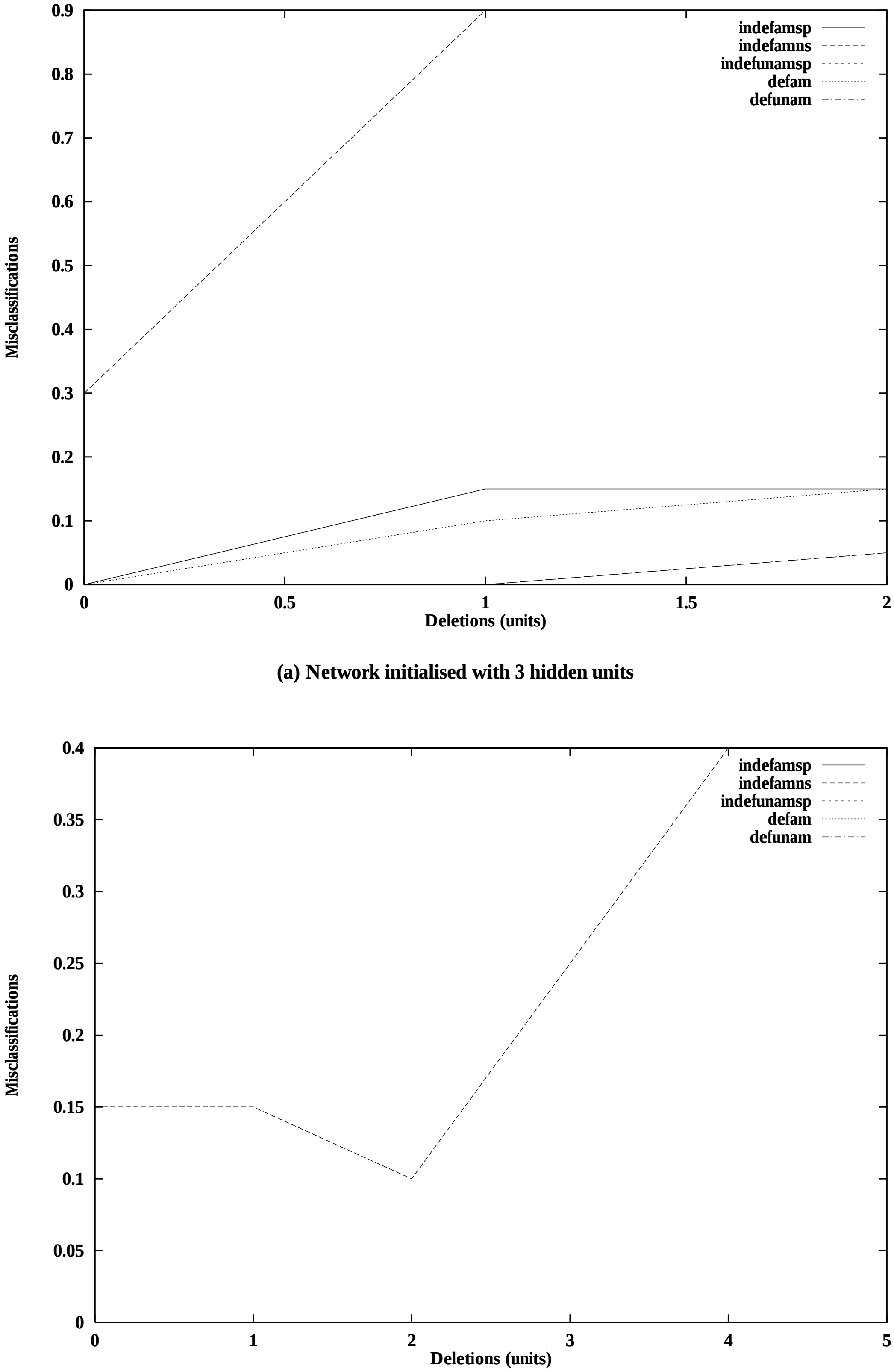
The reason for this difference is made clear by considering a correspondence between network resources and machine functions analogous to that made by Chrisley and Holland (1994) for the SRN's in their study. In terms of Moore machines, the h

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

Chapter



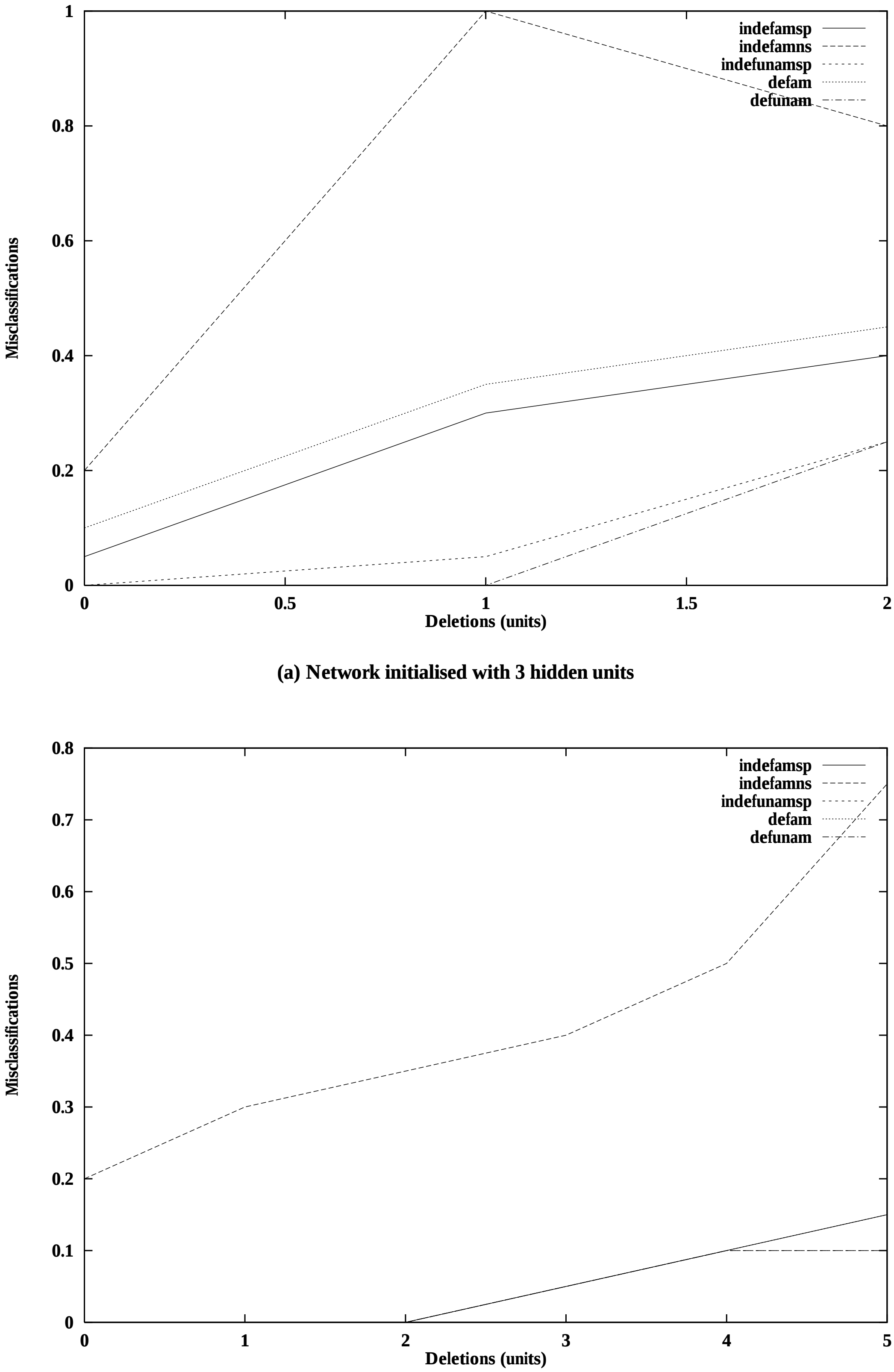
(a) Network initialised with 3 hidden units

(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.

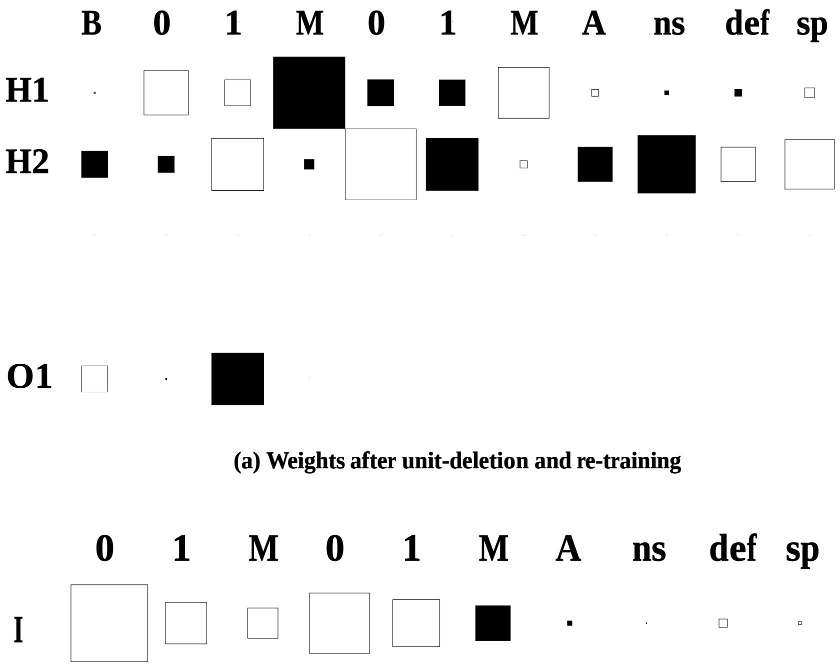
Skeletonisation based on random selection of hidden units **In the interests of investigating the effectiveness of the relevance measure as a means of selecting units for deletion, the above experiments were repeated using random selection of units t**



(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) 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

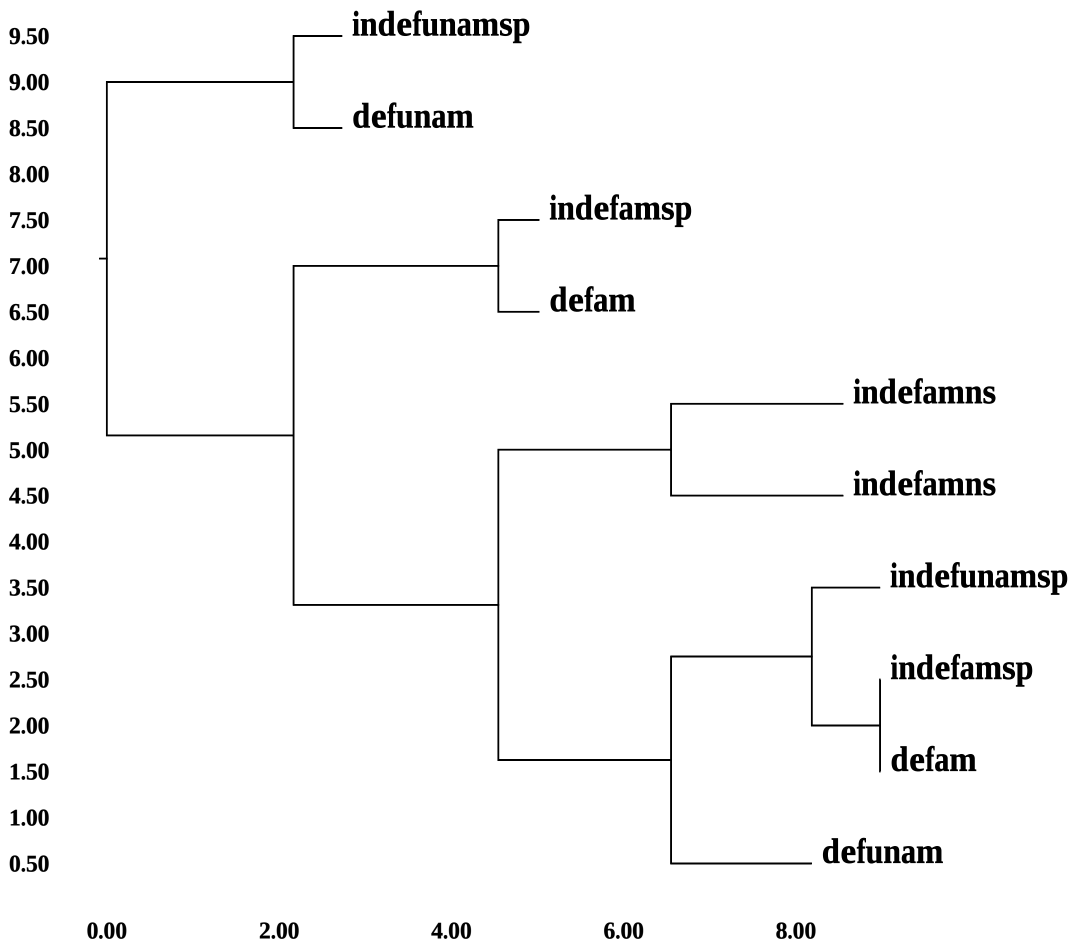


Figure 7.6: Cluster analysis of hidden-unit values after convergence on the ten-pattern data-set. Key: (un)am = (un)ambiguous, ns = non-specific, sp = specific, (in)def = (in)definite

tat st ca ana ys s

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.

í C uster ana ys s

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

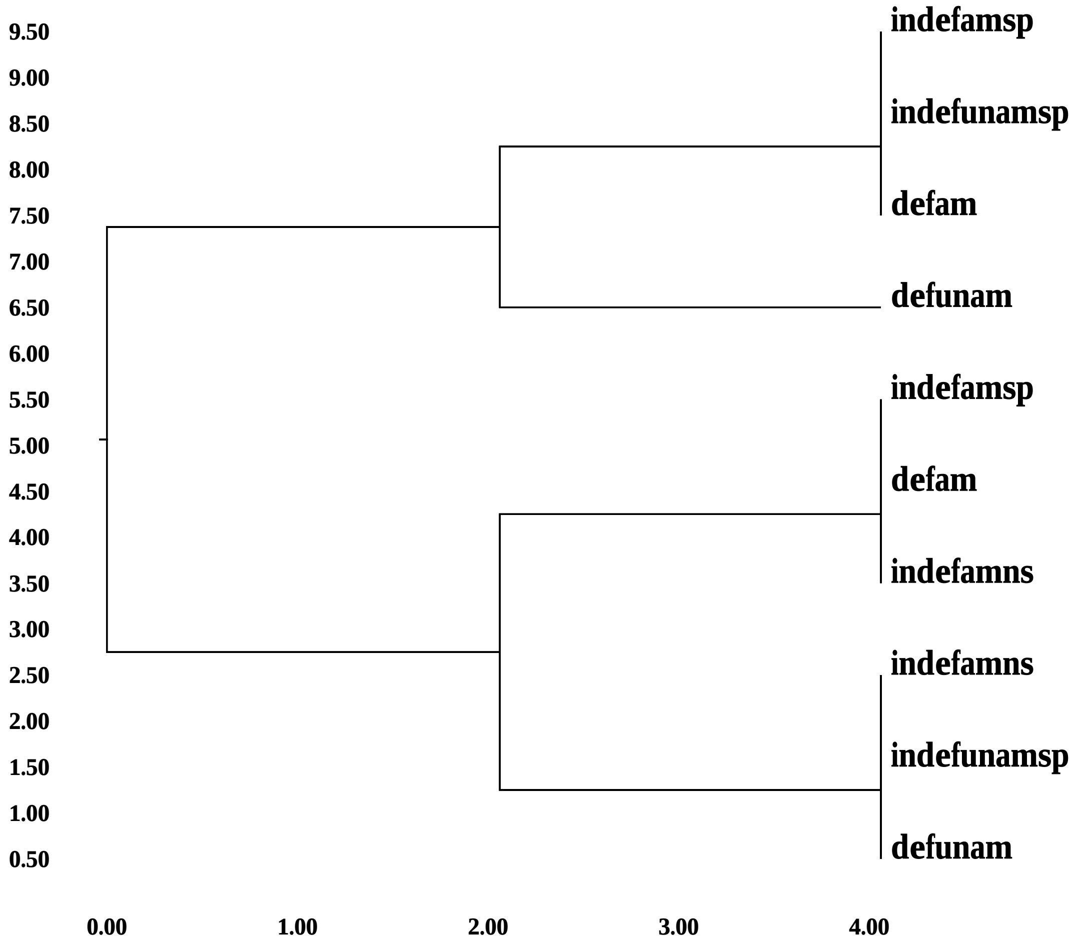


Figure 7.7: Cluster analysis of hidden-unit values after convergence on the ten-pattern data-set after deletion of unit with lowest relevance.

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.

Co par son w t cascade corre at on

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.

u ary

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.

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).

ı —2Mode n sequence earn n us n recurrent cascade corre at on

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, progressin

(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.

—2Cascade corre at on as a ode of representat ona redescr pt on

—i2 Cascade corre at on and t e ode

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

2.2 o es of e e ents of cascade corre at on n ode n redescr pt on

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

est n for

limitedness of the network meant that performance again improved fastest on the most salient feature.

—2Cascade corre at on and s e eton sat on

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.

Co par son w t ot er wor on exp c tat on

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

972(s)2.Q32.48528(i)7.98641(n)]f0.56yn[(s)-348.287(k)4.9616(a)-35.516(a)-3.19(t)-5.76246(m)12.96972832(s)2.394641(n)

iwh76573(84D.27w(o)4.972(f)-268.7672(i)7.98641((e)10.56'4(t)(e)t)-594(u)4.973.19055(s)2.39682(c)1)40.6321(y)40.645551nr.g6

formal constraints on the RR process such as initial configurations and the relative influence of external and internal factors in causing change.

ί 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

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.

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 netwo

Donald, M. (1994). Representation: Ontogenesis and phylog

Gentner, D., Rattermann, M. J., Markman, A., & Kotovsky, L. (

Karmiloff-Smith, A. (1979b). Micro- and macro-developmen

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

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*

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*

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. *Psychologioer5(R)18.5614(e)10.5616(v)-3.19055(i)7.98641(e)10.5616(w)4.795*

Vinter, A., &Perruchet, P. (1994). Is there an implicit leve l of representation? Bel
Brain Sciences, *17*(4), 730–1.