Running Head: A PDP APPROACH TO SEMANTIC COGNITION

A Parallel Distributed Processing approach to semantic cognition: Applications to conceptual development

Timothy T. Rogers
Medical Research Council
Cognition and Brain Sciences Unit
Cambridge, UK

James L. McClelland
Center for the Neural Basis of Cognition
and Department of Psychology
Carnegie Mellon University
Pittsburgh, PA

Abstract

Over the first year of life, infants gain conceptual skills which allow them to construe semantically related items as similar, even when they have few if any directly-perceived attributes in common. Moreover, this skill first encompasses only broad semantic categories, and only later extends to more subtle distinctions, when conceptual and perceptual similarity relations do not coincide. In this paper we suggest that a new mechanism must be added to the mix of possible bases for this observed developmental change. In agreement with many others, we suggest that infants' earliest conceptual representations are organised with respect to certain especially useful or salient properties, regardless of whether such properties can be directly observed. However we suggest that in many cases this salience may itself be acquired, through domain-general learning mechanisms that are sensitive to the high-order coherent covariation of directly-observed stimulus properties across a breadth of experience. To support this argument we will describe simulations with a simple PDP model of semantic memory. When trained with backpropagation to complete queries about the properties of different objects, the model's internal representations differentiate in a coarse-to-fine manner. As a consequence, different sets of properties come to be especially "salient" to the model at different points during development. Such dynamics provide a simple account of the acquisition of conceptual structure from perceptual experience.

A Parallel Distributed Processing approach to semantic cognition: Applications to conceptual development

Although conceptual development in infancy is a wide and varying field of study which encompasses many different points of view, there are, broadly speaking, two sets of empirical findings that are relevant to understanding the competencies that emerge over the first year of life. First, infants as young as 3-4 months are sensitive to visual similarities existing across discriminably different stimulus events, and can capitalize on these to guide their behavior in preferential-looking and dishabituation studies. In some cases, this perceptual learning skill is sufficient for very young infants to discriminate natural semantic categories, on the basis of the visual similarity and variability of the stimulus items encountered during habituation and test (e.g. Eimas & Quinn, 1994). However at the youngest testable ages, infant behaviour in this respect seems to be strongly linked to the visual structure of the specific items encountered during habituation and test, and does not appear to reflect stored knowledge about the conceptual relationships that may exist among stimulus items (Mareschal, French, & Quinn, 2000; Quinn & Johnson, 1997). Thus several researchers have suggested that very young infants are capable of forming “perceptual categories” to guide their expectations in looking tasks.

The second set of findings suggest that, toward the end of the first year of life, infants are less bound to the immediately-observed perceptual properties of objects, and are sometimes able to respond to objects on the basis of somewhat abstract properties that are not directly apparent in the situation. In a series of studies employing several related methods, Mandler and her collaborators (e.g. Mandler & Bauer, 1988; Mandler, Bauer, & McDonough, 1991; Mandler & McDonough, 1993, 1996; Mandler, 1997, 2000) and other investigators (e.g. Pauen, 2002b, 2002a) have argued that pre-verbal children’s choices of replicas of objects to touch, to play with, and to use in imitations of actions performed by others, suggest that at the earliest testable ages (about seven months) infants treat objects that share certain informative but abstract properties (such as self-initiated movement) as similar to one another, even if the objects happen to differ in many other perceptible respects. Moreover, infants at this age seem to be sensitive predominantly to very general conceptual distinctions among perceptually-varying items, and only later come to honor finer distinctions when perceptual and conceptual relations do not happen to coincide (e.g. Mandler & Bauer, 1988; Mandler & McDonough, 1993; Mandler, 2002). There is considerable dispute regarding the interpretation of these data—specifically, whether they can be explained solely with reference to perceptual learning mechanisms; whether or not confounding factors have been adequately controlled; and whether or not the infant behaviour really reveals the first emergence of conceptual knowledge structures. But whichever position one takes on these issues, the data raise two important questions about conceptual development in infancy. First, how do infants come to construe semantically related items as similar to one another when they have few (or perhaps no) directly-perceived properties in common? Second, why are coarse-grained conceptual distinctions available before finer-grained distinctions?

Other investigators have offered answers to one or both of these questions. These answers come with somewhat different slants on the empirical evidence; and both the evidence itself and the theories are the subject of considerable current discussion, as witnessed by the other articles in this volume. Here, we would like to raise the possibility that a process not considered by any of these investigators needs to be added to the mix of possible bases for the characteristics of children's earliest conceptual abilities. Specifically, we will suggest that the learning mechanisms that give rise to infants' earliest concepts are strongly sensitive to patterns of <u>coherent covariation</u> amongst the directly-perceived properties of objects and events. This sensitivity influences the first conceptual distinctions that emerge in infancy, as well as determining which properties of objects infants will weight in particular semantic tasks. Our goal will be first to show that children's initial conceptual distinctions may reflect the effects of coherent covariation; and second, that the importance infants assign to particular cues may reflect the fact that these cues vary coherently with others, and thereby come to <u>acquire</u> salience.

The mechanism that supports these effects derives from our assumptions about the nature of the representations and processes that subserve cognition generally, and semantic cognition in particular: namely, the assumptions inherent in the parallel distributed processing (PDP) framework. Accordingly, we will make our arguments with reference to a simple PDP model of semantic cogntion first described by

Rumelhart (Rumelhart, 1990; Rumelhart & Todd, 1993). This model has strongly influenced our thinking about the acquisition of conceptual knowledge, and we believe that the principles that govern its behaviour can shed light on many different aspects of semantic cognition (Rogers & McClelland, in press). In this paper we will focus on properties of the model that may help to explain how infants are able to represent conceptual similarity relations among perceptually varying stimulus items, and why they are first sensitive to coarse conceptual distinctions. The principles we will emphasize will not and cannot refute the possibility that factors raised by other investigators are also at work. However, if we are able to establish that these principles can play a role in shaping children's conceptual distinctions and in determining what cues they are sensitive to, our account raises questions about the degree to which it is necessary to invoke some of the proposals that have been offered by others.

A Brief Survey of the Landscape

Though most theorists may agree that infant abilities in semantic tasks change in some fashion over the course of the first year of life, there is a diversity of opinion on how best to characterise the underlying cause of the observed differences. It is difficult to do justice to the richness of the literature in this regard; however it will be useful for us to briefly survey some of the stances others have taken on this issue, to set the stage for our later consideration of the PDP approach. Painting with broad strokes, we will here outline four positions regarding the nature of the knowledge that permits older infants to treat items from the same conceptual domain as similar, even if they do not share directly-perceived properties. We should note that these positions need not be considered all mutually exclusive of one another—in some cases they are complimentary, and in other cases not.

Perceptual enrichment. One possibility offered by several researchers (e.g. Quinn & Eimas, 1997; ?; Quinn & Johnson, 1997; Mareschal, 2000) is that older infants come to discern conceptual relationships on the basis of learned associations among directly-perceived properties of items across different episodes and situations. For example, older infants, on the basis of their encounters with particular dogs, cats, birds, and other animals, have learned to associate the various perceptual properties that co-occur together in these encounters—including some properties that may be directly-observable in static photographs and toy replicas (such as eyes, legs, and fur) as well as some that are not (such as patterns of movement and behaviour). Capitalizing on this stored knowledge, older infants in experiments like Mandler's make inferences about properties of the stimulus objects that cannot be directly observed in the experiment, but which have been directly observed on past occasions. On this view, semantically related items come to be associated with similar sets of properties, and as children gain experience they can come to exploit conceptual similarity relations on this basis.

Investigators who adopt this perspective emphasize studies of perceptual categorisation in infancy, which have convincingly demonstrated that by 3-4 months of age infants are sensitive to visual similarities among discriminably different stimulus items, and can employ these similarities to direct their expectations in preferential looking and dishabituation procedures (e.g. Younger & Fearing, 2000; Behl-Chada, 1996; Eimas & Quinn, 1994). For example, when young infants are habituated with a series of cat photographs, habituation generalises to novel cat photos, but not to photographs of dogs (Quinn, Eimas, & Rosenkrantz, 1991). A series of related studies has demonstrated that infant behaviour in this respect depends upon the visual similarity and variability of the specific items viewed during habituation, and it has been noted that it is not necessary to assume they reflect any prior knowledge about the classes to which the stimuli belong (Mareschal et al., 2000). Nevertheless these studies are viewed as being important for understanding conceptual development, because they demonstrate that in many cases, the visual similarities apparent from photographs of real objects can be sufficient to allow the discrimination of semantic categories. The suggestion is that the same perceptual-learning mechanisms observed in such experiments, operating over a much longer interval, may be sufficient to explain the ultimate acquisition of conceptual representations (Quinn & Eimas, 1997, 2000; Quinn, 2002) through day-to-day perceptual experience.

Advocates of the perceptual-enrichment view have tended to focus on understanding the perceptual-learning skills of very young infants, and upon putting forth the argument that within-experiment perceptual learning could be the basis of successful discrimination of both fairly broad semantic categories (such as mammals versus furniture; see Behl-Chada, 1996; Younger & Fearing, 2000; Quinn & Johnson,

1997) and somewhat narrower categories (such as cats versus dogs as cited above), based on the perceptual structure of the items encountered during the experiment (Quinn & Eimas, 2000). There has been less detailed emphasis on understanding how infant behaviour changes with increasing age, and less acceptance of claims by others that performance in various tasks is conceptually rather than perceptually based.

Initial salience of particular perceptual properties.

A related possibility put forward by Rakison and colleagues is that, when infants first appear to demonstrate knowledge of conceptual relationships, they are in fact relying upon certain directly-observed perceptual properties that are inherently salient (Rakison, in press). For example, 12-month-olds may tend to treat toy replicas of animals as similar to one another, and as different from artifacts, because the animals all share legs where most artifacts do not, and legs are especially salient to infants at this age. Specifically, Rakison proposes that at the end of the first year of life, infants assess similarity based mainly or even exclusively on whether or not objects share the same large external parts, and whether they are seen to exhibit the same patterns of movement. Subsequent conceptual development arises from a process of association-based enrichment, as children learn the correlations among these salient properties.

The idea that observed patterns of motion might provide the initial impetus for discriminating animate from inanimate objects was proposed in earlier work by Mandler (1988, 1992); and the hypothesis received some support from Bertenthal's (1993) finding that, even at 3 months of age, infants discriminate mechanical from biological patterns of motion. Rakison has extended the notion that movement patterns form the first basis for conceptual distinctions by raising the possibility that children might notice that items with one kind of large external part (legs) tend to move in one way, whereas items with another kind of large external part (wheels) tend to move in a different way. Such correlations among highly salient properties, which Rakison suggests are shared among category members at a fairly superordinate level (animals have legs and move in one way, vehicles have wheels and move in another), would lead to an early emergence of relatively broad category distinctions; once these sets of salient properties became associated with other properties that are less salient initially, the infant would be able to treat diverse members of a superordinate category the same way, even if none of the highly salient features were actually present in the test stimulus (see Rakison, in press).

Evidence for Rakison's view stems from experiments that follow up on some of Mandler's own work, in which semantic domain (e.g. animal or artifact) is pitted against the external parts possessed by the objects (e.g. whether or not they have legs or wheels). Across a set of such studies, Rakison and colleagues find that 12-month-old infants discriminate wheeled objects from objects with legs, but do not otherwise discriminate animals from artifacts (Rakison & Butterworth, 1998b, 1998a; Rakison & Poulin-Dubois, 2001; Rakison & Cohen, 1999). For example, when confronted with hybrid objects, such as truck bodies with animal legs or cows with wheels, the infants appear to treat artifacts with legs as no different from normal animals; and animals with wheels as no different from wheeled artifacts. The main factor determining their selection of objects to touch or to use for imitation appears to be whether the object has wheels or legs. Thus, Rakison suggests that wheels and legs are among the set of properties that infants find especially salient at 12 months of age. When tested with normal animals and vehicles, infants may appear to be discriminating items on the basis of their semantic relationships, but in fact they are attending only to particular salient properties, according to Rakison (in press).

Rakison also suggests that the inherent salience of external parts and patterns of motion may provide an explanation of why broad semantic distinctions are the first to appear in development. Specifically, the initial salience of certain object properties makes these easier to attend to and learn about than other, less salient properties that happen to differentiate more fine-grained categories. As in the perceptual-enrichment view, infants first discern broad semantic distinctions because they first acquire the associations among properties shared by items in the same semantic domain. However in Rakison's view, the reason these associations are the first to be learned is that such properties are initially and inherently more salient to infants (Rakison, in press).

The interpretation of Rakison's findings is challenged to some degree by data from other studies which suggest that infants can discriminate animate from inanimate objects, even when they share similar large external parts, and are not observed to be moving. For example, Pauen (2002a) found that 11-month-olds reliably discriminate toy animals from toy furniture in an object examination paradigm, even when the replicas are stylised in such a way that items in both categories have similar-looking “legs”, similar textures, and similar salient parts (e.g. the “eyes” on the animals are physically the same as the “knobs” on the furniture, etc.). In fact, infants in this experiment were just as likely to discriminate animals from furniture as was a second group of 11-month-olds who were tested with realistic toy replicas that were perceptually quite different from one another. The results suggest that the conceptual judgments of 11-month-olds are not always influenced by perceptual similarities and differences, and reflect instead an appreciation of what kind of thing the object is.

A key factor contributing to the difference between Rakison’s and Pauen’s results may be that Pauen did not pit two highly salient and informative cues (such as legs and wheels) against other factors that may be weaker. When faced with some items that have wheels and others that have legs, infants may tend to group items on the basis of these features to the exclusion of other informative similarities and differences. However, Pauen’s data demonstrate that 11-month-old infants are nonetheless capable of discriminating animals from artifacts when there are no obvious single features that differentiate the domains—a finding that is difficult to attribute to the inherent salience of large external moving parts. Thus her findings appear to be consistent with the two approaches considered below, in which children rely on their emerging conceptual knowledge at least in some tasks.

Conceptual vs. perceptual knowledge.

According to the approach taken by Mandler (e.g., Mandler, 1990, 1992, 1997, 2000, 2000), the competencies exhibited by older pre-verbal infants reflect a form of knowledge representation which is qualitatively different from that apparent earlier in life. Specifically, Mandler distinguishes between two forms of knowledge possessed by infants: perceptual and conceptual knowledge. Perceptual knowledge encompasses knowledge about what things look like, and on the basis of the dishabituation studies described above, is available to infants by at least 3 months of age. Conceptual knowledge encompasses knowledge about object kinds: it allows infants (as well as older children and adults) to understand that different objects are of the same kind, regardless of whether they are perceptually similar. On this view, the ability of 7- to 9-month-old infants to treat perceptually disparate items as similar to one another (and perceptually similar items from different domains as distinct) provides the earliest evidence for an influence of conceptual knowledge on behaviour. As discussed earlier, Mandler's claims have sometimes been controvertial because of uncertainties about whether perceptual factors could account for some aspects of her results; however the findings of Pauen's (2002 (a and b) recent studies seem to support Mandler's claim that children in the second half of the first year of life show sensitivity to conceptual distinctions in object examination tasks.

According to Mandler, early conceptual representations are built upon representational primitives which she calls image-schemas (Lakoff, 1987). Image-schemas are structures that capture knowledge about relatively abstract spatio-temporal characteristics of objects and events, such as containment, self-initiated movement, and contingency. Mandler differentiates image-schemas from perceptual representations precisely because they capture similarities among inputs that may be superficially quite different. For example, a dog and a bird may move in ways that are perceptually quite distinct; but in both cases the movement is self-initiated. An image-schema for self-initiated movement thus permits infants to represent an element of similarity across these perceptually distinct events. Accordingly, initial concepts are built from such primitives, when infants notice and represent relationships among them. For example, the initial concept animal might be built from image-schemas such as moves-by-itself, moves-irregularly, moves-contingently, interacts-at-a-distance, etc. Infants arrive at the concept animal when they realise that self-moving objects also happen to be those that behave contingently, show irregular patterns of motion, interact with objects at a distance, etc. That is, the initial animal concept results from noticing and representing the relationships amongst these image-schematic primitives. The further ability to discriminate replicas of animals from artifacts, even when these are completely static, similarly arises when infants "notice" that things that move by themselves also tend to have limbs on the bottom, faces, and other directly observable properties. On the basis of such knowledge, infants understand that different self-moving, contingently-interacting objects are of the same kind; and that such items tend to have certain observable external parts. Subsequently, they no longer need to observe an object in motion, or behaving contingently, to "categorise" it as such.

It is not fully clear to us whether Mandler believes that the set of image schemas children first apply are predetermined by innate characteristics of their perceptual or conceptual systems, or whether she believes that they are discovered through the application of some sort of very general purpose mechanism of acquisition. Relatedly it is not clear how early in life the first image schemas are available to differentiate concepts. Mandler believes that the data from infants under about 7 months cannot shed light on these matters, because the looking-tasks with which very young infants are tested only tap perceptual representations, and not conceptual or image-schematic knowledge. Thus it is difficult to know whether young infants possess conceptual knowledge, but do not express it in laboratory tasks; or whether this knowledge first emerges at about 7 months.

In any case, Mandler's view is that by 7-9 months children conceptualize all instances of self-initiated movement as effectively the same, and from this knowledge the early distinction between animate and inanimate things falls out. Thus on this view, the competencies displayed by infants at 7-9 months of age reveal the presence of knowledge structures which, though they may or may not be abstracted from perceptual experience, are qualitatively different from perceptual representations in that they allow the infant to ignore irrelevant perceptual variability and to zero in on the more abstract commonalities and differences that reliably discriminate kinds. Items that have different overall shapes, parts, colors, and textures may be treated as similar to one another if they exhibit the same motion characteristics, or if they share other observable properties that have been incorporated into the infant's image-schemas. Conversely, items with similar outward appearances may be treated differently if they engage different image-schemas. The first image-schemas available to infants describe patterns of motion, and hence the first concepts represented discriminate animate from inanimate objects. Further conceptual development, and the progressive differentiation of concept representations, arises from a continuing process which yields new image-schemas as well as new knowledge about the relationships among image-schemas, although the precise mechanism by which this process occurs is not spelled out.

Emergence of an initial domain theory. A fourth possibility derives from the theory-theory approach to cognition(Carey, 1985; Gopnik & Meltzoff, 1997; Keil, 1989; Gelman & Williams, 1998). On this view, different entities are construed as being the same kind of thing when they are understood to share certain "core" properties—that is, non-observable characteristics that are causal in the sense that they give rise to the item's observable attributes and behaviors. For example, the concept "animal" might include such core properties as agency, rationality, and goal-directedness (e.g. Gergely, Nadasdy, Csibra, & Biro, 1995).

Core properties in the theory-theory tradition serve a function similar to that of Mandler's image-schemas—that is, they permit infants (and older children and adults) to conceive as similar items that may differ in many peripheral respects. If the core properties of animacy include agency and goal-directedness, for example, then any items attributed these characteristics will be understood to be the same kind of thing, regardless of how perceptually dissimilar they may be. However, where Mandler stresses that new image-schemas and conceptual representations may be discovered by the infant through development, theory theorists often emphasise that some important core properties cannot be acquired through experience, and must be specified innately (e.g. Carey & Spelke, 1994, 1996).

In principle this addresses the question of why certain very broad distinctions are apparent early in life—they are based, in this view, on initially available core properties, rather than on less essential properties that must be learned. However this stance still raises questions about how young children happen to know that a given object with particular directly-observable properties also possesses certain crucial but non-observable core properties. One answer (Gelman, 1990; Carey & Spelke, 1994) is that there are initial tendencies to associate certain perceivable properties (e.g., legs) with certain core properties (e.g., agency). Under this position, one can see how the results reported by Rakison might be explained. Specifically, the 12-month-old's judgments would rely upon certain directly-perceived object properties (legs again), which by virtue of an innate mechanism, have been associated with the relevant core properties (e.g. agency). Learning could then allow the infant to associate the non-core properties of objects (such as their surface features, for example) with their core properties, allowing subsequent judgments to be based on non-observable properties.

The Essence of our Argument

The positions we have reviewed offer different answers to what is effectively the same question: On what basis do infants come to conceive of different objects as being of the same kind? For Quinn, Johnson, Mareschal and colleagues, the answer is that conceptually related items come to be associated with similar constellations of attributes on the basis of perceptual learning. For Rakison, conceptually related items share certain inherently salient perceptual properties, including large external parts and patterns of movement. According to Mandler, an understanding of conceptual relations arises from descriptions of objects provided by image-schematic knowledge structures, most of which emerge from a process of perceptual analysis. According to Carey, conceptual relations are determined by non-observable core causal properties, at least some of which cannot be acquired from perceptual learning and must be innate.

To these ideas we would like to add one further suggestion: Perhaps infants treat objects as being of the same kind because they are sensitive to patterns of experienced coherent covariation of properties across objects. This proposal appears to be distinct from the proposals that have been offered by the investigators whose work we have reviewed above. Although many of them discuss the idea that correlational learning plays a role in category formation, none of these accounts really consider anything beyond the role of pair-wise correlations.

Many of the arguments that have been offered by protagonists of other approaches are explicitly or implicitly sensitive to the concern expressed by some proponents of theory theory (e.g. Keil, 1989; Murphy & Medin, 1985; Gelman & Williams, 1998), that correlation-based learning mechanisms are too underconstrained to provide a basis for concept acquisition. The reason is that there are simply far too many spurious or uninformative pair-wise correlations in the world for such correlations to provide a good basis for concept learning. For example, Keil, Carter Smith, Simons, and Levin (1998) point out that although virtually all washing-machines are white in color, being white is not critical to the concept <u>washing-machine</u>. By contrast, all polar bears are white, and in this case, “whiteness” seems to be more important to the concept <u>polar bear</u>. How do people know that the former pairwise correlation (between whiteness and washing-machines) is not particularly important, whereas the latter correlation (between whiteness and polar bears) is?

On the basis of such arguments, investigators who may otherwise have little common ground often agree that infants must begin life with what R. Gelman (Gelman, 1990; Gelman & Williams, 1998) calls enabling constraints—that is, an initial state that constrains to which among the blizzard of correlations yielded up by the environment infants will become sensitive. Thus, for example, Rakison adresses this critique by suggesting that children initially learn about correlations among highly salient properties—effectively giving some correlations a privileged status relative to others. Similarly Mandler promotes image-schemas as providing descriptors that shine through the welter of detail captured directly by perception; and Carey explicitly contends that innate knowledge provides the necessary guidance to bootstrap concept acquisition in the face of an overwhelming amount of irrelevant peceptual information.

What we would like to suggest is that such a bias for the rapid learning of certain correlations over others can itself arise from higher-order patterns of coherent covariation among stimulus properties—so that the properties that first become useful, informative, or salient to infants, and that are easiest to associate with one another, are just those that participate in the strongest patterns of coherent covariation across many different events and situations. In the remainder of the paper, we will focus on demonstrating how this can be, by considering a simple computational model of semantic memory.

The Rumelhart Model

The model we will focus on was first put forward by Rumelhart (Rumelhart, 1990; Rumelhart & Todd, 1993), to demonstrate that the semantic content captured by propositional spreading-activation models such as Collins and Quillian's (1969) could also be coded in the distributed representations acquired by a feed-forward PDP network trained with backpropagation. To that end, they constructed a network whose architecture reflects the structure of a simple proposition, as shown in Figure 1. The first term of the proposition is coded with local representations in the Item layer; different relations are coded with local representations in the Relation layer; and the various completions of a given proposition are represented by individual units in the layer labelled Attribute. When presented with a particular Item and Relation pair in the input, the network must turn on the attribute units in the output that correctly complete the proposition. For example, when the units corresponding to robin and can are activated in the input, the network must learn to activate the output units move, grow and fly. The particular items, relations, and attributes used by Rumelhart and Todd (1993) were taken directly from the corpus of propositions used by Quillian (1968) (indeed, though we have left them out in our work, Rumelhart and Todd included item input units for general as well as specific concepts in their model). Hence, when the network had learned to correctly complete all of the propositions, it had encoded the same information stored in the propositional spreading-activation model.

Insert Figure 1 about here

The network consists of a series of nonlinear processing units, organized into layers, and connected in a feed-forward manner as shown in the illustration. Patterns are presented by activating one unit in each of the item and relation layers, and allowing activation to spread forward through the network, subject to the constraints imposed by the weights and the sigmoid activation function of the units. In order to perform correctly, the network must find a configuration of weights that will produce the correct states across output units for a given pair of inputs—when it has done so, it can be said to "know" the domain (insofar, of course, as this is reflected in the training materials).

To find an appropriate set of weights, the model is trained with backpropagation (Rumelhart, Hinton, & Williams, 1986). First, an item and relation are presented to the network, and activation is propagated forward to the output units. The observed output states are then compared to the desired values, and the difference is converted to a measure of error. The partial derivative of the error with respect to each weight in the network is computed in a backward pass, and the weights are adjusted by a small amount to reduce the discrepancy.

Although the model's inputs are localist, each individual Item unit projects to all of the units in the layer labelled Representation. The activation of a single item in the model's input, then, generates a distributed pattern of activity across these units. The weights connecting item and representation units evolve during learning, so the pattern of activity generated across the Representation units for a given item is a learned internal representation of the item. Though the model's input and target states are constrained to locally represent particular items, attributes, and relations, the learning process allows it to derive distributed internal representations that do not have this localist character.

Because each item is represented locally in the the input, the model is initially given no information about how the objects in its virtual world are related to one another. For example, the pine and oak inputs are no more similar to one another than each is to the salmon input. Due to the small, random values of the weights from these units to the Representation units, the patterns initially all have very similar internal representations, with only sight random differences. However, as the model learns to complete particular propositions, these representations gradually change, and as we will see later, items with different attributes come to be differentiated, while items with many shared attributes continue to be represented by similar patterns of activity across the Representation units (Rumelhart & Todd, 1993). Thus, the connections that link Item and Representation units, once the model has been trained, can be viewed as encoding a set of semantic similarity relationships among a set of otherwise arbitrary markers. The connections in the rest of the network then reflect the learned mappings between these internal representations, in combination with input from the relation context, and explicit object properties.

An important departure from other representational schemes (including those used in some other connectionist approaches) is that the internal representations acquired by the Rumelhart network are not semantic features or in any other way directly interpretable semantically. Individual units in the model's Representation layer do not encode the presence or absence of explicit, intuitive object properties. Rather, these distributed representations are abstracted from the featural decomposition of objects represented in the output layer. The network's representations capture the similarities existing among different kinds of objects, not the actual semantic properties themselves. These can be activated from the combined effects of the units in the distributed representation, working in concert with units in other parts of the network.

The similarities that the network acquires through learning provide a basis for semantic generalisation and induction: items with similar internal representations will tend to generate similar outputs. To demonstrate this, Rumelhart and Todd (1993) trained a network like the one shown in Figure 1 to correctly complete propositions about a variety of plants and animals. When it had learned the correct responses for the items and relations in its environment, they examined its generalization behavior using a procedure often employed in the connectionist literature (see, for example, Miikkulainen, 1993). They added a new input node to represent a novel kind of bird (a sparrow), and used the backpropagation of error to assign a representation to the item based solely on the information that it was a kind of bird. Specifically, they turned on the sparrow and isa units in the input, and propagated activity forward to the output units. They then calculated error on just the "Bird" output unit, and used backpropagation to find an appropriate pattern of activity for the sparrow. Once a suitable representation had been found, they stored the pattern by adjusting only the weights connecting the sparrow unit to the Representation units. Using this procedure, the model found a representation for the new object that allowed it to activate the output property "Bird" with the isa relation, given the weights forward from the Representation units that encode its knowledge of the entire domain. Thus, the network's ability to assemble an appropriate representation relied upon the information stored in its weights.

Once they had established a representation for sparrow based on the information that a sparrow is a bird, Rumelhart and Todd (1993) queried it about the sparrow in all four relation contexts. Although the network had derived its internal representation solely from the information that the sparrow is a kind of bird, it strongly activated all the properties shared by the familiar birds: has wings, has feathers, can grow, can fly, isa animal, etc. However, properties unique to individual birds (such as is red or can sing) were not strongly activated. Examining the model's internal representations, the authors found that the network had come to represent the sparrow with a pattern of activity similar (but not identical) to those generated by the robin and canary inputs. This similarity led the model to attribute to the sparrow the properties shared by the robin and canary, but not their idiosyncratic properties.

The Rumelhart model, together with a range of related work, (McClelland & Rumelhart, 1986; Rumelhart et al., 1986; McClelland, McNaughton, & O'Reilly, 1995; McClelland, St. John, & Taraban, 1989), suggest a general framework for understanding semantic cognition that is quite different from propositional theories, and from other theories that assume a mechanism of categorisation as the vehicle for semantic knowledge storage, generalisation, and retrieval. According to this view, semantic representations of objects consist of patterns of activity across a set of units in a connectionist network, with semantically related objects represented by similar patterns of activity. In a given semantic task, these representations may be constrained both by incoming information about the object (in the form of a verbal description, a visual image, or other sensory information) and by the context in which the semantic task is performed. In turn, the instantiation of the representation in unit activation states allows the system to correctly complete the semantic task. On this view, all semantic knowledge is stored in and processed by the same set of hardware elements (the weights and units respectively). Generalization of stored knowledge to new items and situations results as a natural consequence of the similarities among object representations in a given context, and not from the operation of a categorization mechanism, or by inference across a stored system of propositions.

Obviously, the Rumelhart model's behavior in this respect depends on the state of its weight matrix at any given point during learning. The accumulation of small weight changes in the network as it learns leads its internal representations of objects to evolve in interesting ways, with consequences for the network's ability to perform various semantic tasks at different points throughout training. In the simulations to come, we will show that the gradual weight changes that occur when the network is exposed to a set of propositions about a subdomain of conceptual knowledge, the resulting progressive differentiation of its internal representations, and the consequent change over time in its generalization and induction behavior, together suggest a mechanism that may help to explain how infants acquire conceptual representations on the basis of perceptual experience.

Progressive Differentiation and Feature Selection in the Rumelhart Model

We will begin by considering how internal representations in the Rumelhart model change over time as a consequence of learning the Collins and Quillian corpus of propositions. This aspect of the model's behaviour was previously considered in a simulation reported in McClelland et al. (1995), and other investigators interested in differentiation of perceptual and conceptual representations in infancy have also presented similar simulations (Quinn & Johnson, 1997; Miikkulainen & Dyer, 1991; Schyns, 1991). Here we replicate the progressive differentiation simulation of McClelland et al. (1995) as a starting point, then go on to consider the implications of this pattern for issues raised in the introduction.

We trained the network shown in Figure 1 with the same corpus of propositions used by Rumelhart and Todd (1993). The weights were initialized to small random values selected from a uniform distribution with a mean of zero and variance of 0.5. With each pattern presentation, a single unit was activated in each of the Item and Relation layers, and weights were adjusted by a small amount (learning rate = 0.1) to reduce the sum-squared-error across output units. The network processed each input-relation pair once in each training epoch, but the order of patterns within an epoch was randomized. In this simulation, the simplest version of the back-propagation algorithm was used; the model was trained without noise, weight decay, or momentum. Weights were updated after every epoch, based on weight error derivatives calculated after processing each pattern. To ensure that the model's output responses relied entirely on input from other units in the network, we assigned to all units in the model a fixed, untrainable bias of -2. Thus, in the absence of input, each unit's state would rest at approximately 0.19, slightly below its midrange. We trained the network for 1500 epochs, at which point each output unit was within 0.05 of its target activation (0 or 1) on every pattern.

Progressive differentiation of item representations. To see how internal representations develop in the network, we stopped training at different points during learning and stepped through the eight items, recording the states of the representation units for each.

Insert Figure 2 about here

In Figure 2 we show the activations of the representation units for each of the eight item inputs at three points in learning. Each pattern of activation at each time point is shown using eight bars, with each representing the activation of one of the representation units. Initially, and even after 50 epochs of training as shown, the patterns representing the items are all very similar, with activations hovering around 0.5. At epoch 100, the patterns corresponding to various animal instances are similar to one another, but are distinct from the plants. At epoch 150, items from the same intermediate cluster, such as rose and daisy, have similar but distinguishable patterns, and are now easily differentiated from their nearest neighbors (e.g. pine and oak). Thus, each item has a unique representation, but semantic relations are preserved in the similarity structure across representations.

Insert Figure 3 about here

The arrangement and grouping of the representations shown in Figure 3 reflects the similarity structure among the internal representations, as determined by a hierarchical clustering analysis using Euclidean distance as the measure of similarity between patterns. At 50 epochs the tree is very flat and any similarity structure revealed in the plot is weak and random. By epoch 100 the clustering analysis reveals that the network has differentiated plants from animals: all the plants are grouped under one node, while all the animals are grouped under another. At this point, more fine-grained structure is not yet clear. For example, oak is grouped with rose, indicating that these representations are more similar to one another than is oak to pine. By epoch 150, the network has learned the correct similarities, and we can see that the learned distributed representations fully capture the hierarchical relations among the input items. The degree of semantic relatedness among the 8 items is reflected by the degree of similarity in the patterns of activity that represent each item when the model has finished learning. Moreover, the model first discovers fairly broad conceptual distinctions, and only later comes to represent more subtle ones.

Insert Figure 4 about here

In order to better visualize the process of conceptual differentiation that takes place in this model, we performed a multidimensional scaling of the internal representations for all items at 10 different points during training. Specifically, the Representation layer activation vector for each item at each point in time was treated as a vector in an 8-dimensional space. The Euclidean distances between all vectors at all points over development were calculated. Each vector was then assigned a 2-d coordinate, such that the pairwise distances in the 2-d space were as similar as possible to the distances in the original 8-d space.

The solution is plotted in Figure 4. The lines trace the trajectory of each item throughout learning in the 2-dimensional compression of the representation state space. The labelled end points of the lines represent the final learned internal representations after 1500 epochs of training. The figure shows that the items, which initially are bunched together in the middle of the space, soon divide into two global clusters (plant or animal) based on animacy. Next, the global categories split into smaller intermediate clusters, and finally the individual items are pulled apart.

Discussion of Differentiation Results. Our simulation replicates results described previously by others (McClelland et al., 1995; Quinn & Johnson, 1997) showing that when a back-propagation network is trained on a set of training patterns with a hierarchical similarity structure, it will exhibit a pattern of progressive differentiation. One interesting aspect of this process is the tendency for the different levels to differentiate in relatively discrete stages, first completing differentiation of at the most general level before progressing to successively more fine-grained levels of differentiation. This tendency to exhibit stage-like learning is a feature of connectionist models that has been considered extensively elsewhere (McClelland, 1989; Plunkett & Sinha, 1992; McClelland, 1994). Our present task is to try to provide the reader with a mechanistic understanding of the progressive differentiation process, drawing on insights expressed in the papers just cited (see also McClelland & Rumelhart, 1988), to explain how stage-like progressive differentiation works in the present case.

With the training set used here, very early in learning, the network comes to represent all the animals as similar to one another, and as quite distinct from the plants. Only later does it come to learn to differentiate the patterns at an intermediate level, and only after that does it learn to differentiate the items from each other at the subordinate level. Why does this occur? To begin to gain an intuitive understanding of this, let us consider how the network learns about the following four objects: the oak, the pine, the daisy, and the salmon. Early in learning, when the weights are small and random, all of these inputs produce a similar meaningless pattern of activity throughout the network. Since oaks and pines share many output properties, this pattern results in a similar error signal for the two items, and the weights leaving the oak and pine units move in similar directions. Because the salmon shares few properties with the oak and pine, the same initial pattern of output activations produces a different error signal, and the weights leaving the salmon input unit move in a different direction. What about the daisy? It shares more properties with the oak and the pine than it does with the salmon or any of the other animals, and so it tends to move in a similar direction as the other plants. Similarly, the rose tends to be pushed in the same direction as all of the other plants, and the other animals tend to be pushed in the same direction as the salmon. As a consequence, on the next pass, the pattern of activity across the representation units will remain similar for all the plants, but will tend to differ between the plants and the animals.

This explanation captures part of what is going on in the early stages of learning in the model, but does not fully explain why there is such a strong tendency to learn the superordinate structure first. Why is it that so little intermediate level information is acquired until after the superordinate level information? Put another way, why don't the points in similarity space for different items move in straight lines toward their final locations? Several factors appear to be at work, but one is key:

> For items with similar representations, coherent covariation of properties across these items tends to move connections coherently in the same direction, while idiosyncratic variation tends to move weights in opposing directions that cancel each other out.

To see how this happens in the model, lets consider the fact that the animals all share some properties (e.g., they all can move, they all have skin, they are all called animals). Early in training, all the animals have the same representation. When this is so, if the weights going forward from the representation layer "work" to capture these shared properties for one of the animals, they must simultaneously work to capture them for all of the others. Similarly, any weight change that is made to capture the shared properties for one of the items will produce the same benefit in capturing these properties for all of the other items: If the representations of all of the items are the same, then changes applied to the forward-projecting weights for one of the items will affect all of the others items equally, and so the changes made when processing each individual item will tend to cumulate with those made in processing the others. On the other hand, weight changes made to capture a property of an item that is not shared by others with the same representation will tend to be detrimental for the other items, and when these other items are processed the changes will actually be reversed. For example, two of the animals (canary and robin) can fly but not swim, and the other two (the salmon and the sunfish) can swim but not fly. If the four animals all have the same representation, what is right for half of the animals is wrong for the other half, and the weight changes across different patterns will tend to cancel each other out. The consequence is that:

> Properties shared by items with similar representations will be learned faster than the properties that differentiate such items.

The preceding paragraph considers the effects of coherent covariation in the weights forward from the representation later in the Rumelhart network. What about the weights from the input units to the representation layer? As previously stated, items with similar outputs will have their representations pushed in the same direction, while items with dissimilar outputs will have their representations pushed in different directions. The question remaining is why the dissimilarity between, say, the fish and the birds does not push the representations apart very much from the very beginning. The answer is somewhat complex, but understanding it is crucial, since it is fundamental to understanding the progressive nature of the differentiation process.

The key to this question lies in understanding that the magnitude of the changes made to the representation weights depends on the extent to which this will reduce the error at the output level. The extent to which change in the representation weights will reduce the error at the output in turn depends on whether the forward weights from the representation layer to the output are able to make use of any changes in the activations of the representation units. Their ability to make use of such changes depends on them already being at least partially organized to do so. Put in other words, we can point out a further very important aspect of the way the model learns:

> Error back-propagates much more strongly through weights that are already structured to perform useful forward-mappings.

We can illustrate this by observing the error signal propagated back to the representation units for the canary item, from three different kinds of output units: those that reliably discriminate plants from animals (such as can move and has roots), those that reliably discriminate birds from fish (such as can fly and has gills), and those that differentiate the canary from the robin (such as is red and can sing). In Figure 5, we show the mean error reaching the Representation layer throughout training, across each of these types of output unit when the model is given the canary (middle plot). We graph this alongside measures of the distance between the two bird representations, between the birds and the fish, and between the animals and the plants (bottom plot); and also alongside of measures of activation of the output units for sing, fly and move (top plot). We can see that there comes a point at which the network is beginning to differentiate the plants and the animals, and is beginning to activate move correctly for all of the animals. At this time the average error information from output properties like can move is producing a much stronger signal than the average error information from properties like can fly or can sing. As a consequence, the information that the canary can move is contributing much more strongly to changing the representation weights than is the information that the canary can fly and sing. Put differently, the knowledge that the canary can move is more "important" for determining how it should be represented than the information that it can fly and sing, at this stage of learning. (The error signal for move eventually dies out as the correct activation reaches asymptote, since there is no longer any error signal to propagate once the model has learned to produce the correct activation).

Insert Figure 5 about here

The overall situation can be summarised as follows. Initially the network is assigning virtually the same representation to all of the items. At this point, what is coherent is only what is shared across everything — the is living, can grow, and isa living thing outputs. All other output properties have their effects on the forward weights almost completely cancelled out. However, because the plants have several properties that none of the animals have and vice-versa, weak error signals from each of these properties begin to cumulate, eventually driving the representations of plants and animals apart. At this point, the common animal representation can begin to drive the activation of output units that are common to all animals, and the common plant representation ban begin to drive activation of outputs common to all plants. The weights so structured in turn allow these coherently-varying properties to exert much stronger influences on the representation units than those exerted by the properties that differ between the birds and the fish. The result is that the individual animal representations stay similar to one another, and are rapidly propelled away from the individual plant representations. Very gradually, however, the weak signals back-propagated from properties that reliably discriminate birds from fish begin to cumulate, and cause the representations of these sub-groups to differentiate slightly, thereby providing a basis for exploiting this coherent covariation in the forward weights. This process eventually propagates all the way down to the subordinate level, so that idiosyncratic properties of individual items are eventually mastered by the net. In short, there is a kind of symbiosis of the weights into and out of the representation units, such that both sets are sensitive to successive waves of coherent covariation among output properties. Each “wave” of properties only becomes coherent after the prior stage of differentiation has occurred.

Adapting the Rumelhart Framework to Accommodate Preverbal Learning

With the above background in place, we are ready to turn to the specific details of our use of the Rumelhart model as a vehicle for illustrating the key points of our argument. To launch our discussion of these issues, consider the following abstract formulation of the Rumelhart model architecture. Here we envision that the two parts of the input represent a perceived object (perhaps foregrounded for some reason to be in the focus of attention) and a context provided by other information available together with the perceived object. Perhaps the situation is analogous to one in which a young child is looking at a robin on a branch of a tree, and, as a cat approaches, sees it suddenly fly away. The object and the situation together provide a context in which it would be possible for an experienced observer to anticipate that the robin will fly away; and the observation that it does would provide input allowing a less experienced observer to develop such an anticipation. Conceptually speaking, this is how we see learning occuring in preverbal conceptual development. An object and a situation or context afford the basis for implicit predictions (which may initially be null or weak), and observed events then provide the basis for adjusting the connection weights underlying these predictions, thereby allowing the experience to drive change in both underlying representations and predictions of observable outcomes.

With this scenario in front of us, we can consider the wide range of different contexts in which the child might encounter an object. Some such contexts will be ones in which the child is watching the object and observing what others might do with it (pick it up, eat it, use it to sweep the floor, etc); another (as in the example above) might be one in which the child is simply observing the object itself, watching the things that it does in various different situations. Several contexts will be ones in which someone engages the child in language-related interactions concerning the object. Some such encounters may involve naming, as when an adult points to an object and says to a child “Look, Sally, it’s a bunny rabbit.” Others may include indicating for the child the various parts of the object (“OK, Sally, let’s pat the Bunny’s tail. Can you see the tail? Here it is!”). Each encounter with a particular object in a given context will give rise to certain observed consequences; and we suggest that the child learns to assign a conceptual representation to each object based on the consequences observed in different situations. The contexts or situations include linguistic ones as well as non-linguistic situations in which the child observes the object either alone or in interaction with other objects. We suggest that conceptual development arises from the learning that occurs across many such situations.[1]

We can now consider how our modeling framework allows us to capture aspects of this learning process, and in particular how useful conceptual representations can be acquired on the basis of such learning. The presentation of the “object” corresponds to the activation of the appropriate pattern of activity over the input units in the Rumelhart model; the context can be represented via the activation of an appropriate pattern over the context units; the child's expectations about the outcome of the event may be equated with the model's outputs; and the presentation of the actual observed outcome is analogous to the presentation of the target for the output units in the network.

On this view, the environment provides both the input that characterizes a situation as well as the information about the outcome that then drives the process of learning. In the example above, the item input corresponds to the visual appearance of the object, and the context input provides the additional source of information that constrains the child's predictions about what will happen next, which take the form of a pattern of activation across the output units. The weights and representations that mediate between the inputs and the outputs constitute the state of knowledge that allows the system to anticipate the outcome of the event, by activating the units that correspond to the predicted conclusion; but the environment contributes again by providing the actual observed event outcome, thereby yielding information the system can use to determine whether its predictions were accurate or not. This outcome information will consist sometimes of verbal, sometimes of non-verbal information, and in general is construed as information filtered through perceptual systems, no different in any essential way from the information that drives the Item and Context units in the network. What this means is that this information is provided by non-conceptual (perceptual, motor feedback, etc) systems and serves as input that drives the learning that results in the formation of conceptual representations. It will of course be obvious that this is a drastic simplification of perceptual, motoric, attentional, and other processes, but we believe the resulting model is useful in that it brings out some of aspects of the processes that may underlie conceptual development.

We can also see that there is a natural analog in the model for the distinction drawn between the perceptual information available from an item in a given situation, and the conceptual representations that are derived from this information. Specifically, the model's input, context, and targets code the "perceptual" information that is available from the environment in a given episode; and the intermediating units in the Representation and Hidden layers correspond to the "conceptual" representations that allow the semantic system to accurately perform semantic tasks.

Simulating Conceptual Differentiation in Infancy

Insert Figure 6 about here

To explore these ideas more fully, we conducted simulations with an alternative implementation of the Rumelhart model, designed to illustrate how forces of coherent covariation can lead to the model to represent semantically related items as similar, even when they have no directly-perceived properties in common, and share only a few unobserved and abstract characteristics.

The alternative implementation is depicted in Figure 6. It differs from the model shown earlier in two respects. First, we have expanded the training corpus to accommodate a somewhat broader range of items. Specifically, we added two new birds, fish, flowers and trees to the set of items from the original network, and an additional set of five different land animals. Thus there are 21 items in total in the new implementation: 13 animals from three categories (birds, land animals, and fish), and 8 plants from two categories (trees and flowers). To allow differentiation of the individual items, and to capture the properties of the land animals, several additional property units were required, as indicated in the Figure. In this case we construe these properties not as verbal propositions but as directly observable properties of objects in the environment. Additionally, because these simulations are meant to address preverbal conceptual learning, the model was not trained to name objects in either simulation—they were exposed to only the is, can, and has patterns. Though we believe that experience with spoken language may play some role in concept acquisition prior to the infant's ability to produce speech, it will be useful to see that the progressive differentiation of representations in the model does not depend upon its being trained to name objects, or to explicitly categorize them in other ways that rigidly express taxonomic relations.

While we have extended the training corpus, the reader will note that we have not gone all the way to producing a fully realistic training environment. First of all, there are many more kinds of things in the world that are not included in this training corpus, and indeed many more examples of each of the kinds that we are actually using. Second, though the properties are in some sense correctly assigned to the items, it is also true that the set of properties is far from complete, and it is not at all clear that those properties that are included are necessarily the most salient, informative, or likely to be available from the environment. As one example, in retaining the properties from the Rumelhart training set we retain the property has skin, which was originally used in (Collins & Quillian, 1969). However, this property is likely not to be as salient or available as, for example, the missing property has eyes. Other properties that some researchers have indicated as being important to discriminating animate from inanimate objects—contingent movement, action at a distance, wheels, surface textures, etc.—are not represented. Many readers may wonder what insights can come from a model based on such inherently incomplete and even somewhat inaccurate training data. Our response to this question is as follows.

The fundamental force that drives learning in our network is not the particular set of properties we have chosen, but the patterns of co-variation that occur among the properties used in the model's training environment. To see this, imagine that we completely relabelled all of the input and output units with completely arbitrary symbols, such as I1 - I21 (for the 21 different items) R1-4 (for the four different relation types) and A1 - A34 for the 34 different attributes. None of this would have the slightest consequence for the process of learning in the network, which depends only upon the degree of overlap in the properties shared across the range of items. Thus, we may agree that has skin is not really a salient feature of animals, but we might also agree that animals do nevertheless share some other salient attribute (e.g. has eyes). Again, the crucial point is that it is not the identity of the properties themselves but their patterns of co-variation that is essential to the model's behaviour.

Insert Figure 7 about here

In Figure 7 we show the degree to which different items in the corpus tend to share the same output attributes, assessed across all contexts. It is apparent that the individual trees all have very similar sets of attributes, as do the flowers, the birds, the fish, and the land animals; that there is considerable similarity between the trees and the flowers and among the three types of animals; and there is very little similarity between the various plants on the one hand and the various animals on the other. In the model, it is the structure present in this similarity matrix, rather than the particular sets of item, relation, and attribute labels used, that govern the models' learning and performance.

The second departure from the original Rumelhart model is that the new implementation was provided with distributed "perceptual" input representations, rather than localist proposition-like inputs. In place of the Item units employed in the original Rumelhart implementation (with one unit standing for each item), we instead construe the input units as representing the subset of perceptual attributes that are apparent from an item's visual appearance — for example, features such as red, big and legs. A particular item is presented to the model by instantiating a pattern of activation across this set of "perceptual" input units. The model might be shown a picture of a robin, for instance, by activating input units corresponding to red, legs, and wings. Each item in the training corpus is represented with a different pattern of activity across input features, rather than by the activation of a single input unit. The extent of feature overlap in these input representations provides a model analog of "visual" similarity in the input. The instantiation of a particular input pattern gives rise to a pattern of activity across Representation and Hidden units, which correspond to internal "conceptual" representations just as in the localist implementation.

We emphasize that the particular attributes that are included in the training corpus were not chosen with the goal of addressing infancy findings per se. Rather, we have employed this expanded corpus to investigate a wide range of dufferent issues related to semantic cognition (Rogers & McClelland, in press). Accepting that the particular perceptual properties to which infants are sensitive may not be precisely those that are expressed by the labelled attributes in the model, recall that it is the pattern of property covariation across items in the training corpus that determines the model's behaviour. The units in the network could be relabelled to better express the particular kinds of perceptual information to which real infants are actually exposed, but of course this would not influence the behaviour of the model. Addition or deletion of properties, alternations of the assignments of particular properties to particular objects, and manipulations of salience could alter the results, but only insofar as these alterations removed certain key properties of the training corpus, namely the presence of the particular pattern of coherent covariation of properties that gives rise to the item similarity structure seen in Figure 7.

<u>Choosing input representations</u>. In generating “visual” input patterns for the 21 items in the distributed-inputs implementation, one is immediately faced with the question of which attributes to include in the input. We have argued that, in principle, all of the properties included in the training corpus are potentially observable, at least in certain circumstances — for example, the property <u>can move</u> can be considered “perceptual” information available from the input whenever one is directly observing a moving object. In other situations, this information is not available directly from perception, but must be inferred — for example, when one observes a stationary cat. Ought the property to be coded then in the input, the output, or both? In fact we don’t think there is a categorical distinction between those properties that are available as the basis for making preditions, and those that should be predicted. In recurrent networks, where a given unit can be at once both an input and an output unit, all attributes can serve both as inputs and as targets (see for example Rogers et al., in press). For simplicity we’ve stayed with the feed-forward architecture, however, so we must adopt a specific policy on the allocation of attributes to input and output.

The policy chosen reflects our primary aim in the current simulation, which is to investigate how the model comes to represent semantically related items as similar, even when they have few or no directly-perceived properties in common. To this end, we employed as “perceptual” input attributes the complete set of <u>is</u> properties from the training corpus (<u>big</u>, <u>pretty</u>, <u>green</u>, <u>red</u>, <u>yellow</u>, <u>white</u>, and <u>twirly</u>) and a subset of the <u>has</u> properties (<u>wings</u>, <u>legs</u>, <u>gills</u>, <u>petals</u>, and <u>branches</u>). Note that these properties are also included as output attributes, along with the remaining <u>has</u> properties and the <u>can</u> properties for all 21 instances.

This choice accomplished three ends. First, the set of attributes chosen were intended to correspond at least approximately to those that might be directly observable in almost any encounter with the item or a picture of it, and thus could provide a simple analog to the visible properties of real-world objects that are likely to be seen regardless of the task or situation. We do not intend to suggest by this that the specific attributes we have chosen correspond to the actual visual attributes of real objects. We simply note that, of the complete set of perceptual properties that one could potential observe in an object, a relatively small subset are likely to be frequently visible (for example, outside parts and colours), whereas others are likely to be perceived only in certain limited circumstances (for example, inside parts or particular behaviours). In the model, the complete range of potentially observable attributes is coded in the model's output attributes, and the subset of attributes that are treated as observable in all contexts and situations are coded in the distributed input.

Insert Figure 8 about here

Second, to the extent that the set of attributes used are the ones that are available in a picture or a scale replica of the object, they provide patterns that can be used as inputs in simulation analogs of experimental tests performed on children, which commonly use pictures or scale replicas as stimuli.

Third, and of central importance for the key points we wish to make, the "perceptual" similarities expressed by this subset of attributes fail to specify any superordinate similarity structure beyond the intermediate category level. As shown in Figure 8, items from the same intermediate categories (birds, fish, mammals, flowers and trees) are "perceptually" similar to one another, but no global similarity between the trees and flowers, or between the birds, fish, and mammals, is available in this subset of properties. Thus the "perceptual" similarities captured by our choice of input patterns provides no impetus for the distributed-inputs simulation to develop a superordinate distinction between plants on the one hand and animals on the other. Yet as we shall see the model does still learn to differentiate the inputs on this basis, since the superordinate structure is still present in the set of target attributes.

Simulation Details. The network was trained as described previously with each pattern appearing once in every training epoch, without momentum or weight decay, and with output units assigned a fixed, untrainable bias weight of -2. However we adopted the following slight changes to the training procedure in this simulation. First, a small amount of noise selected from a Gaussian distribution with a mean of zero and a variance of 0.05 was injected into the inputs of all the hidden units throughout training. Second, the model was trained with a learning rate of 0.005 instead of 0.1; with the larger training corpus and the noisy hidden units a larger learning rate occasionally prevented the model from completely mastering the training corpus. Third, weights were updated after every ten pattern presentations (rather than at the end of each epoch). The order in which patterns were presented to the network was determined randomly in each epoch. Finally, because the model's internal representations in the early epochs of training can be influenced by the particular configuration of random weights with which it is initialized, the results we describe here are averaged across five network runs trained with different random starting weights.

Differentiation of Representations for Familiar Items

Insert Figure 9 about here

To see how the model's internal "conceptual" representations for the familiar items in the training corpus change over development, we stopped training after every 125 epochs, and stepped through all 21 items, recording the patterns of activation across Representation units. Figure 9 shows hierarchical cluster plots of the distances among representations at three different points during learning, averaged across the five training runs. Again, items differentiated in a coarse-to-fine manner.

Note that at epoch 1000, the model behaves in a manner analogous to the 7-to-9-month-old infants described in Mandler's (2000) experiments. That is, it "groups together" items that share the few properties that reliably discriminate broad semantic domains — properties such as can move, has skin, has roots and has leaves. The model may appear as though it is "using" these attributes as its basis for grouping items, even though none of these useful properties is coded in the input. In fact, there are no properties shared by all plants or all animals represented in the input; but nevertheless the model first differentiates its internal representations with respect to this global semantic distinction. Thus the weights that project from the Input to the Representation layer effectively serve the function that Mandler (2000) attributes to image-schemas, and that Carey (2000) attributes to core casual properties: they allow the model to "group together" a set of items that have disparate visual appearances, on the basis of a few abstract shared properties that are not directly represented in the input.

The model shows a progressive pattern of differentiation for the same reason outlined earlier. Initially all items are represented as similar to one another by virtue of the intital small random weights, and the model reduces error by learning to activate output units shared by all items in the environment (e.g. can grow). However plants and animals, because they have few properties in common in the output, generate quite different error signals, and these gradually lead the model to distinguish them slightly. As the plants and animals slowly come to receive different internal representations, a kind of "feedback" loop arises in the learning mechanism: the weights projecting forward from the Representation layer are able to use the dimension of variability that separates plants and animals to reduce error on the sets of output properties shared across items in each domain. Consequently the error derivatives that come back from these properties to the Representation units grow increasingly large, propelling the plant and animal represenations further apart. This in turn gives the forward weights even more information to work with, and they adjust to captialize still further on the growing distinction represented between plants and animals.

Over the course of learning, the constellation of output properties that reliably discriminate plants from animals effectively become more "salient" to the model, in the sense that only these properties generate a coherent error signal across the training corpus in the early epochs of learning. That is, the property can move is not initially predisposed to be any more salient or informative than any other attribute; but as the model begins to learn, the coherent covariation between can move and other attributes that reliably discriminate animate from inanimate objects in the training environment lead these properties to dominate learning in the model in the early epochs of training. As a consequence, these properties more strongly constrain the model's internal representations — leading the model to discover an early weight configuration that "filters" input similarities in such a way that differences between items in the same domain are minimized, and differences between items in different domains are emphasized.

As the network masters properties that vary coherently across domains, and as the smaller weight changes that accumulate across other output units very gradually allow the model to differentiate more fine-grained categories such as bird and fish, the dynamics shift to favour of learning about the properties that vary coherently with these intermediate categories. That is, as the bird and fish representations pull apart from one another, properties such as can fly and can swim begin to produce coherent error signals that produce large derivatives and therefore large weight changes, at which point this learning comes to strongly influence representational change in the model. The consequence of these dynamics is that different sets of properties dominate the learning that drives representational change in the model at different points during training.

Simulating the Object-Examination Experiment

With these ideas in hand, we can begin to see how the model provides a basis for understanding the phenomena reviewed in the introduction: the ability of older infants to treat perceptually disparate stimuli as similar to one another on the basis of properties not present in the input available at the time of test. The explanation offered by model is similar in some respects to that suggested by other investigators: properties that reliably discriminate broad semantic domains are more important in determining the model's behavior early in development than are other properties, and items that share these properties are represented as conceptually similar, even if they differ in many other respects. What sets our account apart from that offered by others is the basis for the importance of these properties. In the case of our model, the salience of properties like movement emerges as a consequence of the sensitivity of the learning mechanism to coherent covariation.

To provide a concrete illustration of how these phenomena in the model might explain data from infant experiments, we conducted a simulation of the object-examination task conducted by Mandler and McDonough (1993). In this experiment, infants were allowed to play with a series of toy objects belonging to the same semantic category. After habituation, they were presented with two new test objects in succession—first, a novel item from the same semantic category, and second, a novel item from a different category—and the authors measured the amount of time the infant spent examining each. If the infants construe the same-category item as novel, they should spend more time examining it, relative to the last trial of habituation. If they further construe the different-category item as novel, they should spend still more time examining it, relative to the last trial of habituation.

Three different experimental conditions are of particular interest in this experiment. First, the authors habituated 9- and 11-month-olds to items from the same broad semantic domain (e.g. a series of animals, or a series of vehicles), which had broadly different overall shapes. In this case, infants of both ages failed to dishabituate to the same-category item, but successfully dishabituated to the different-category item—indicating that they could “group together” items from the same general category, even though they were perceptually fairly dissimilar, and could construe these as different from an out-of-category item. In the second condition, the authors habituated 9- and 11-month-olds to a set of items from the same intermediate category (e.g. a series of dogs), and then tested them with an item from a different category, but within the same broad domain (e.g. a fish). In this case, the different-category test item had a fairly different shape and some different parts from the habituation items; but nevertheless, 9-month-olds did not dishabituate to this test item—indicating that they did not construe the dogs and fish as different, despite their perceptual differences. In contrast, 11-month-olds dishabituated to the out-of-category item, indicating successful discrimination of the different kinds of animal. Finally in the 3rd condition, the authors habituated 9- and 11-month-olds to a series of toy birds modelled with their wings outspread, and then tested them with a novel bird (also with wings outspread) or with a toy plane. Even though all habituation and test items had a similar overall shape, infants at both ages dishabituated to the different-category item, indicating successful discrimination of the categories despite perceptual similarities. Thus in sum, infants at 9 months discriminated items from broadly different semantic categories, both when the habituation and test items had variable shapes and parts, and when they had a grossly similar overall shape; but they failed to discriminate dogs from fish, despite perceptual differences among these items. 11-month-olds, in contrast, discriminated both the broader and more specific categories in all conditions.

We find the Mandler and MacDonough (1993) findings (and the similar findings of Pauen, 2002b) to be of particular interest because they indicate an early ability to discriminate broad conceptual domains for both perceptually similar and perceptually disparate items; and a developmental change between the ages of 9 and 11 months in the ability to differentiate sub-categories within the same conceptual domain (i.e. dogs and birds). Infants at both ages were tested with the same stimulus items—hence the different patterns of behaviour cannot be explained solely with reference to the perceptual structure of the stimulus materials themselves. While it remains possible to attribute some aspects of the findings to perceptual rather than semantic similarities, the developmental change indicates differences in the representations and/or processing by infants at different ages—changes that are consistent with the developmental processes that operate in the Rumelhart model.

In the model analog of the experiment, we "habituate" the model with novel stimulus items that are perceptually similar to one another, and that belong one of the four of the categories with which the model is familiar (birds, fish, trees and flowers). We then test the model with novel test items that include a semantically-related item from the same category that has few attributes in common with the habituation items, and which is therefore "perceptually dissimilar"; and a semantically distinct item from a different category that shares many perceptual attributes with the habituation items (and is therefore "perceptually similar" to these). We then measure the similarities among the model's internal representations to determine which of the test items the model construes as "novel" with respect to the habituation items, at different stages of learning.

Representing test items in the distributed implementation. To present the model with a novel item, we simply apply a previously unseen pattern of activity across the input units, which corresponds to the directly-observed visual properties of the novel item. We can then inspect the resulting pattern of activity across Representation units to determine what "conceptual" representations the model assigns to the item. To simulate the experiment, we needed to create input patterns that allow us to manipulate the "visual similarity"—the degree of overlap in directly-perceived features—of the habituation and test items. Specifically, we created a set of habituation items that had many input attributes in common, and which belong to the same category; a semantically-related test item that belongs to the same category but which has few properties overlapping with the habituation items; and a semantically distinct test item which shares many perceptual properties with the habituation stimuli, but which belongs to a different category.

Input patterns with these properties are shown in Figure 10. There are four "categories" of items represented. In each category, items 1-4 share many perceptible attributes in common. The fifth item in each case has few directly-perceived attributes in common with its category neighbours; but in all cases it shares one especially useful and informative property with them. For example, bird-5 shares the property wings with the other birds, but otherwise has no perceptible attribute in common with them.

Insert Figure 10 about here

For each of the four categories, there is one "perceptually similar" out-of-category item from the contrasting domain, and one from the contrasting category in the same domain. For example, trees 1-4 have many properties that overlap with bird-5, and many that overlap with flower-5. In this sense, both bird-5 and flower-5 are "perceptually" more similar to trees 1-4 than is tree-5. This construction allows us to pit perceptual feature overlap in the input against semantic relatedness in our analog of the preference task: we "habituate" the network with four similar items from the same category (e.g. trees 1-4), and then "test" it with a perceptually dissimilar item from the same category (e.g. tree-5) and a perceptually similar item from the contrasting category (flower-5) or domain (bird-5).

Insert Figure 11 about here

Figure 11 shows the similarities among some of the habituation and test items that are apparent from the overlap in attributes shown in Figure 10. In each case the "habituation" items are those labelled 1-4 in the tree. As is apparent from the figure, the fifth category member is less similar to the four habituation item than is the test item from the contrasting category or domain. Using this set of patterns we can investigate changes in the model's ability to discriminate both broad and more specific conceptual categories as it learns.

Habituation and test procedure. To simulate the familiarisation-preference procedure, we present the model with four familiarisation items (e.g. birds 1-4), a semantically-related item (e.g. bird-5), and a semantically-distinct item (e.g. fish-5 or tree-5) in sequence, and record the internal representations it derives for each. We then calculate the distance between the centroid of the representations for the 4 habituation items, and each of the two test items. To map from these distances to an analog of the infant behaviour, we adopt the following assumption: we assume that the model construes as more novel, and therefore tends to choose, whichever test item gives rise to a representation that has the largest distance from the centroid of the habituation items; and that the likelihood of choosing one object over another increases with the discrepancy in their respective distances from the habituation centroid. Consistent with this assumption, we use the following formula to determine the likelihood of choosing the target item given the relevant mean distances:

$$p_b = s\sigma(d_b - d_w);.$$

In this formula, p_b is the probability of choosing the between category test item, σ is the logistic function, d_b is the mean distance from the between-category test to the four habituation items, d_w is the mean distance from the within-category test item to the four habituation items, and s is a scaling parameter that determines the degree to which the model is sensitive to the difference $d_b - d_w$. The probability of choosing the within category test item, p_w, is just $1 - p_b$. Intuitively, the equation indicates that when the between and within-category items are equally distant from the habituation items, the model is equally likely to choose to manipulate either; but as one test item gets further from these relative to the other, the likelihood of choosing it for examination increases.

Simulation details. The model was trained just as described in the previous section. Training was halted every 125 epochs, the weights were stored and the simulation of the habituation experiment was begun. For each of the 20 test items shown in Figure 10, the model generated an internal representation, simply by presenting the appropriate pattern across the input units and observing the subsequent pattern of activity arising across Representation units.

To test the network's tendency to choose a between-category test item, at both the global and intermediate (i.e. basic) levels, the distances among its internal representations of habituation and test items were calculated. Specifically, the centroid was determined for each set of four habituation items (birds 1-4, fish 1-4, flowers 1-4 or trees 1-4), and the distance between this centroid and each of the corresponding test items was calculated. These distances were then entered into the formula shown above to determine the probability that the model "preferred" the different-category item. Each category (e.g., the bird category) had one perceptually dissimilar same-category test item (e.g., bird 5); one perceptually similar global different-category test item (e.g., flower 1 or tree 4); and one perceptually similar, intermediate-level different-category test item, from the contrasting category in the same domain (e.g., fish 5). This yielded four comparisons testing intermediate category differentiation, and four testing global category differentiation. The simulation was run five times with different starting weights, yielding 20 data points in total for each level of contrast. The data we report are averaged across these trials.

Note that the model was never trained with the items used in the habituation procedure, and no weight changes were made during the habituation process. The representations the model generates for these items simply reflect knowledge the model has accumulated on the basis of learning in the normal training environment.

Results

Insert Figure 12 about here

Figure 12 shows the similarities among the representations the model generates for one set of habituation and test items (the bird category), at three different points throughout learning. During the very early epochs of training all items are represented as quite similar to one another, although the organization of items does reflect to some extent the degree of overlap in the perceptual properties from which the internal representations are derived. For instance, bird-5 is represented as less similar to the other birds than either flower-1 or fish-5. The reason is that at this point, there is no information accumulated in the network's weights that allows it to map reliably between parts of the representation space and any of the directly-perceived attributes that generate the error signal that the network is using to find internal representations. Items with similar sets of directly-observed attributes thus generate similar, weak error derivatives at the Representation units, and the network finds similar internal represenations for these, regardless of their semantic relationships to one another. However as the model learns more about the coherent covariation among properties in the extended training corpus, this picture begins to change: the model first differentiates the global different-category test item (the flower) from the five birds, and later comes to differentiate the intermediate-level different-category item (the fish) from the birds—even though both of these test items share more "perceptual" properties with birds 1-4 than does bird 5.

Figure 13 shows the proportion of times the model chose the different-category test item over the same-category test item throughout early training, for between-domain test items (e.g. tree or flower vs. bird, a global discrimination) or for intermediate-level between-category test items (e.g. tree vs. flower or bird vs. fish). Early in learning, all items are represented as similar to one another, so same- and different-category test items (at both the global and intermediate levels) are equally distant from the habituation centroid. Consequently, the likelihood of choosing the between-category item is near chance. As training proceeds, this likelihood begins to change — the model first begins to reliably choose the different-category test item from the contrasting global domain, and later begins to choose the local between-category test item. Like infants, the model's ability to differentiate items on conceptual grounds emerges first for broad semantic distinctions, and only later for more specific ones.

Insert Figure 13 about here

Finally, it is worth noting that, in generating its internal representations of the novel items after learning, the model seems to lend special weight or importance to the properties that covary coherently together in its training environment. This is reflected in the fact that the model continues to group the same-category item with the habituation items through learning—even though it shares very few of its properties with these. For instance, of the 4 features that describe bird5 (shown in Figure 10), only one is consistently shared by the four other birds seen during habituation. By contrast, the test item fish5 shares 5 of its 6 properties with at least half of the birds seen during habituation. Considering just the overlap in features, then, fish5 is more similar to the habituation items than is bird5 (as is clear from Figure 11).

However, the one property that bird5 does have in common with the four habituation birds—wings—happens to covary coherently with many other properties in the model training corpus. By contrast, the 5 properties that fish5 shares with the habituation birds are all somewhat idiosyncratic in the training corpus, and do not covary coherently together with anything else. As a consequence of the dynamics of learning described earlier, the coherently covarying property comes to contribute more strongly to representational change than do the various idiosyncratic properties—so that the network treats as similar those items that share the coherent property, even if they do not have many other attributes in common; and treats as different items that differ for the coherent property, even if they share many idiosyncratic attributes. In this sense, the model's sensitivity to coherent covariation leads it to treat some properties as more "important" for semantic representation than others—that is, these properties come to have an acquired salience for the model.

Discussion

There are several points we would like to make on the basis of the preceding simulations. To begin, one of our aims has been to support the idea that the the Rumelhart model, as simple as it is, might nevertheless provide some insight into the acquisition of semantic knowledge from different kinds of perceptual experience across a range of events and situations; and might specifically have relevance to understanding aspects of the early conceptual development of preverbal infants. On the present construal of the model, both the input and output attributes can be viewed as coding various aspects of similarity and difference among objects encountered in the environment, as detected by perceptual processes across different events and situations. In any given situation, such perceptual similarities may not yield much information about which objects should be treated as the same kind of thing. However, repeated encounters with a range of objects across a variety of contexts leads to a state of knowledge in which some attributes exert particularly strong constraints on the semantic system's internal representations, allowing it to treat items that share these properties as similar even when they differ in many other respects.

The key factor contributing to this state of knowledge is the influence of coherent covariation on the learning processes that govern weight changes in the system. The ease with which the system can learn that a particular item has a certain property depends upon the degree to which the property is shared by other items with similar internal representations. However, the similarities represented by the model at any point in time themselves depend upon the mappings it has learned between internal representations and particular output properties. At a particular time the model treats as salient those properties that covary coherently within the clusters of objects that it has learned to differentiate at that time. As we have seen, such properties are easiest for the system to acquire, and thus dominate the representational change that occurs as the system learns. Once these properties have been mastered, representational change slows dramatically until, on the basis of minute weight changes accumulating from non-coherent properties, the system finds a new organisation of its internal representations which renders a new set of properties coherent. Such properties become easier to learn and propel new, rapid representational change until they are mastered. Thus, coherent covariation among stimulus attributes at different granularities spurs successive waves of differentiation, with different stimulus properties acquiring salience at different points during development. It is this changing sensitivity to patterns of coherent covariation that we propose to add to the repertoire of possible mechanisms that may contribute to the process of conceptual development.

It is important to acknowledge a key point that must be treated as an essential presupposition of our analysis. The mechanism we have described depends upon the following proposition being true:

> The conceptual distinctions to which children are first sensitive are just the ones that reflect the strongest patterns of coherent covariation present in the child's perceptual experience.

We anticipate this presupposition will be viewed by our critics as revealing a fatal flaw in our reasoning. To the contrary, we would like to suggest the assumption in itself does not preclude a role for other contributing factors to concept development, including additional assumptions similar to those that other investigators have offered. In other words, we believe that more than one of the alternative positions can be partially correct; and indeed we believe it is likely that all of them have at least some partial validity.

This validity stems in part from the fact that the child's perceptual experience reflects, not just the structure present in the world, but the ways in which this structure is filtered by the child's perceptual system.

Consider, first, the suggestion of Rakison that some kinds of perceptual information may be more salient than others. This idea is a very old one, and is almost certain to have some validity. For example it is clear that the brain has specialized mechanisms for motion detection (e.g. Zeki, 1978), and that motion strongly engages attention. The training materials used in our modelling efforts could not be justified without accepting that they presuppose a selection of certain properties for inclusion, and that the basis for inclusion reflects assumptions about the availability and salience of information in the input. That is, the model's behavior depends upon the particular training patterns to which it is exposed, and these training patterns incorporate implicit or explicit assumptions about which properties of objects are available or salient in the input, and which are not.[2]

Consider, second, the suggestion of Mandler that conceptual knowledge emerges from a process that generates conceptual descriptors, which represent certain stimulus events as similar despite differences in perceptual details. For example, as we have seen, different instances of self-initiated movement, in Mandler's view, come to be treated as conceptually the same, even though they may be quite distinct perceptually: a flounder and a gazelle seem to move in very different ways, but somehow the fact that they are both self-propelled becomes apparent to infants, and provides the basis for forming the concept animal. For Mandler, the problem of understanding how infants arrive at this realization is solved by supposing that there exists a process that yields up descriptions of the flounder's movement and the gazelle's movement that have something in common. Our analysis, like Mandler's, also depends upon the assumption that different kinds of animal movement share some common representational element. We include a unit in the model that corresponds to the attribute can move—an attribute shared by all animals and none of the plants. To use such a unit in our simulations is essentially to specify that all forms of animal movement overlap in some perceptually-detectable way, thereby providing a basis for seeing them as having something in common. Similar points can be made about all of the attributes in our training set.

It is important to realize, however, that in imbuing the network with these perceptual skills, we have not predisposed it to assign special weight or salience to some attributes rather than others. The network must still discover the category structure inherent in its inputs, and must still determine which of the attributes are “important” for organising its internal representations. To see this, consider what happens when the Rumelhart network is exposed to three different events — a red robin, flying; a white goat, walking; and a red rose, growing. The target representations we have used “build in” similarities and differences between experiences with these different events. The move attribute unit in the model encodes a degree of similarity between the goat’s walking and the robin’s flying, while the walk and fly units code a degree of difference between these as well. The red unit codes an aspect of similarity between the robin and the rose, which differentiate both from the goat. All of these elements of similarity and difference are coded in target patterns provided for the model’s attribute units. However, feedback from the environment does not indicate which of them the model should use to “group together” the different items in its internal conceptual representations. In fact, on the basis of their movement patterns and colors, there is no perceptual basis provided by the environment for determining which two items are of the same “kind” in this example. What this will depend on, instead, is the fact that moving co-varies coherently with other properties across all items in the training corpus, whereas being red does not—thus the learning process will issue a greater salience to movement.

At issue is the young infant’s (and the model’s) capacity to “choose” which of the many different kinds of detectable similarities and differences among stimulus objects should be used to determine which items are of the same kind. While the model “builds in” an initial ability to detect various elements of similarity and difference between experiences (and could be further augmented to reflect differential salience as previously noted), there is nothing in the initial state as we have specified it for these simulations that inherently lends more weight or importance to (for example) the move attribute relative to others. Hence the model has no “built in” basis to represent as similar two items that share the capacity to move than two items that share the attribute is red. The competencies exhibited by infants at 9 months of age in the studies described above — their ability to zero in on such properties as self-initiated movement, or movement-enabling parts such as legs and wheels, and to employ these as the basis for representing objects as similar to one another — are not given to the network in its initial state.

There remains, in our view, a fundamental unresolved question: To what extent does our use of units to detect aspects of similarity across various different kinds of events amount to building in domain-specific knowledge? This is where we anticipate that the opinions of other investigators will differ, with Carey, Spelke and some of the other theory-theorists lining up behind the view that it does, and others whose perspectives we have considered perhaps leaning toward the view that it does not. We would like to adopt a somewhat agnostic position on this point, and simply contend that children’s perceptual processing systems can act as “filters” that influence the degree to which distinct events will be perceived as similar or different. We resist the assertion that different perceptual filtering systems are brought to bear on different “kinds” of things. To be sure, different “kinds” of things draw upon on different types of information at some point fairly early in the life-span (e.g., movement may be important for some things, shape and color for others); and different types of information may be filtered in different ways from birth. We do not deny that such filtering can influence the knowledge infants acquire about different kinds of things. However we do not accept the idea that perceptual filtering systems are wired up in advance to apply different filters to the very same type of information, depending on what “kind” of object or event is being processed by the filter.

The principles illustrated by the Rumelhart model do not and cannot refute other claims about the factors that may potentially contribute to conceptual development. However, accepting that the mechanism we have identified may be a contributing factor, our simulations have implications for each of the other viewpoints we reviewed earlier.

In agreement with advocates of perceptual enrichment, the PDP framework suggests that conceptual knowledge acquisition is spurred by domain-general learning that is based on perceptual experience. Our further point here is that the sensitivity of correlation-based learning mechanisms to coherent covariation among stimulus properties, and the resultant influence on acquired feature salience, provides a previously unrecognized mechanism by which such domain-general perceptual learning can give rise to internal representations that capture similarity structure different from that which is available directly from the perceptual input provided by test stimuli.

Rakison’s claim that certain directly-observed properties are especially salient to 12-month-old infants is not inconsistent with our theory; and indeed, nothing in our approach refutes the possibility that some attributes are initially and inherently more salient to infants than others. However, our simulations also demonstrate that coherent covariation can lead certain properties to have an acquired salience. Thus, the empirical demonstration that infants in semantic tasks are most sensitive to large, external parts, or to patterns of motion, need not reflect a learning-independent perceptual salience of these properties—this salience might emerge as a consequence of domain-general learning over the first year of life. If so, this might explain why 12-month-olds are inclined to emphasise external parts such as legs or wheels as a basis for grouping objects (Rakison & Poulin-Dubois, 2001). Note that, under this account, there is nothing special about external parts; any properties that varied coherently across domains could potentially provide the basis for differentiating concepts, including overall shape and structural characteristics that are likely to covary coherently with other properties.

We also suggest that the PDP framework has similarities with some aspects of Mandler's ideas about the emergence of conceptual representations from perceptual experience. The state of knowledge captured by the network at a fairly early point is similar in many respects to that attributed by Mandler to 7- to 9-month-old infants. That is, the model treats as similar perceptually varying items that happen to share characteristics such as self-initiated movement that are not directly available in stimuli used in experiments. For example, our model first represents animals as similar to one another and as distinct from plants, despite the fact that there are no input properties held in common between the birds and the fish, and many input and target properties that differentiate these items. One might argue that the model's "conceptual" and "perceptual" representations capture different information about the similarities among objects, as Mandler suggests—immediately available perceptual similarities are captured in the input, and conceptual similarities are captured in the internal representations. Moreover, the learning process captured by our model provides a new mechanistic basis for the extraction of conceptual representations from perceptual experience, different from but not inconsistent with the conceptual discovery process that Mandler attributes to infants.

Finally, Carey (2000) suggests that the conceptual distinctions made by infants early in life cannot be acquired and must reflect initial domain-specific knowledge about non-obvious "core" conceptual attributes.

We have addressed some of Carey's ideas in detail in related work (see Rogers & McClelland, in press), but for the time being we should point out that there is one sense in which, on our view, infants are biologically prepared to acquire concepts. The influence of coherent covariation on concept development depends upon the infant's initial ability to detect elements of similarity and difference among particular events. Without such a facility, there can be no basis for patterns of covariation to influence knowledge acquisition. The theory thus assumes that, among the many various elements of similarity to which infants are initially sensitive, there exist some that vary coherently across semantic domains. Our simulations suggest that it may not be necessary to attribute to infants initial domain-specific predispositions to lend special salience, weight, or attention to specific core properties, since this salience may emerge as a consequence of coherent covariation.

We would like to close with a final thought on what will perhaps be viewed as the most difficult contention of our theory: the assumption that "perceptual" input contains elements that remain invariant from one event to another, despite discriminable differences in the particular instantiations of such common elements across different events. This assumption amounts to the claim that, while there may be differences, say, between the specific legs of one animal and another, there is also at least some perceptually-given element of similarity. Our network assigns internal representations to objects on the basis of coherent covariation tabulated across such presumed elements of similarity, and the results described here may suggest to some readers that the model depends strictly on this literal overlap for its success in discovering the category structure of the domain.

Although our simulations do not illustrate it, other work has shown that PDP networks can sometimes discover patterns of coherent co-variation among items that have no direct feature overlap of any kind. Such a situation is illustrated in the simulations of Hinton (1986, 1989), in a network with an architecture similar to the Rumelhart model. Hinton trained his network with a corpus of information about the relationships among individuals in two different families, one English and one Italian. When given an individual's name and a relationship as input (e.g. <u>Joe</u> and <u>father</u>), the model was trained to activate the name of the appropriate individual in the output (e.g. Joe's father Bill). Each person was represented with a localist unit in the model, both at the input and the output level, so there was no direct overlap of any kind for the various individuals in the two families. Even so the network discovered internal representations that captured the position of each individual within the respective family tree—for example, assigning near-identical internal representations to the English and Italian "grandchildren"; near identical representations to the English and Italian "uncles," etc. Although there was no direct overlap in the input and output representations for, say, the English and Italian grandfathers, the network's internal representations discovered the higher-order commonalities across the two different families—that is, it discovered that both grandfathers entered into a similar set of relationships with others in their respective families, and on this basis came to represent the individuals as similar. This and other work (Altmann & Dienes, 1999, e.g.) is consistent with the idea that coherent covariation can be discovered by networks even in the absence of any overlap of the input and output patterns of activation.

While direct feature-overlap for different items may not be necessary for coherent covariation to exert an influence on the acquisition of internal representation, we believe such overlap is in fact made available by our perceptual systems, at least for some co-varying attributes of objects if not for all of them. In either case, the fundamental point remains the same: the ability of networks to exploit such covariation plays a crucial role in the early emergence of semantic categorization abilities, and in the global-to-local differentiation of such categories as a gradual result of experience, beginning in infancy and continuing throughout life.

References

Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. Science, 284, 875. (With reply by G. Marcus)

Behl-Chada, G. (1996). Basic-level and superordinate-like categorical representations in infancy. Cognition, 60, 105-141.

Bertenthal, B. (1993). Infants' perception of biomechanical motions: Instrinsic image and knowledge-based constraints. In C. Grandrud (Ed.), Visual perception and cognition in infancy (p. 175-214). Hillsdale, NJ: Erlbaum.

Carey, S. (1985). Conceptual change in childhood. Cambridge, MA: MIT Press.

Carey, S. (2000). The origin of concepts. Cognition and Development, 1, 37-42.

Carey, S., & Spelke, E. (1994). Mapping the mind: Domain specificity in cognition and culture. In L. A. Hirschfeld & S. Gelman (Eds.), (p. 169-200). New York, NY: Cambridge University Press.

Carey, S., & Spelke, E. (1996). Science and core knowledge. Journal of Philosophy of Science, 63, 515-533.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240-247.

Eimas, P. D., & Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. Child-Development, 65(3), 903-917.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate/inanimate distinction as examples. Cognitive Science, 14, 79-106.

Gelman, R., & Williams, E. M. (1998). Enabling constraints for cognitive development and learning: A domain-specific epigenetic theory. In D. Kuhn & R. Siegler (Eds.), Handbook of child psychology, Volume II: Cognition, perception and development (Vol. 2, 5 ed., p. 575-630). New York: John Wiley and Sons.

Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. Cognition, 56, 165-193.

Gopnik, A., & Meltzoff, A. N. (1997). Words, thoughts, and theories. Cambridge, MA: MIT Press.

Hinton, G. E. (1986). Learning distributed representations of concepts. In Proceedings of the 8th Annual Conference of the Cognitive Science Society (p. 1-12). Hillsdale, NJ: Erlbaum.

Hinton, G. E. (1989). Learning distributed representations of concepts. In R. G. M. Morris (Ed.), Parallel distributed processing: Implications for psychology and neurobiology (p. 46-61). Oxford, England: Clarendon Press.

Keil, F. (1989). Concepts, kinds, and cognitive development. Cambridge, MA: MIT Press.

Keil, F., Carter Smith, W., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. Cognition, 65(2-3), 103-135.

Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. Chicago, IL: University of Chicago Press.

Mandler, J. (2000). What global-before-basic trend? Commentary on perceptually based approaches to early categorization. Infancy, 1(1), 99-110.

Mandler, J. (2002). On the foundations of the semantic system. In E. M. Forde & G. Humphreys (Eds.), Category specificity in mind and brain. East Sussex, U.K.: Psychology Press. (To appear in:)

Mandler, J. M. (1988). How to build a baby: On the development of an accessible representational system. Cognitive Development, 3, 113-136.

Mandler, J. M. (1990). From perception to conception. In P. van Geert & L. Mos (Eds.), Developmental psychology. New York: Plenum.

Mandler, J. M. (1992). How to build a baby II: Conceptual primitives. Psychological Review, 99(4), 587-604.

Mandler, J. M. (1997). Representation. In D. Kuhn & R. Siegler (Eds.), Cognition, perception, and language (Vol. 2). New York, NY: Wiley.

Mandler, J. M. (2000). Perceptual and conceptual processes in infancy. Journal of Cognition and Development, 1, 3-36.

Mandler, J. M., & Bauer, P. J. (1988). The cradle of categorization: Is the basic level basic? Cognitive Development, 3, 247-264.

Mandler, J. M., Bauer, P. J., & McDonough, L. (1991). Separating the sheep from the goats: Differentiating global categories. Cognitive Psychology, 23, 263-298.

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. Cognitive Development, 8, 291-318.

Mandler, J. M., & McDonough, L. (1996). Drinking and driving don't mix: Inductive generalization in infancy. Cognition, 59, 307-355.

Mareschal, D. (2000). Infant object knowledge: Current trends and controversies. Trends in Cognitive Science, 4, 408-416.

Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. Developmental Psychology, 36(5), 635-645.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. G. M. Morris (Ed.), Parallel distributed processing: Implications for psychology and neurobiology (p. 8-45). New York: Oxford University Press.

McClelland, J. L. (1994). Learning the general but not the specific. Current Biology, 4, 357-358.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102, 419-457.

McClelland, J. L., & Rumelhart, D. E. (1986). A distributed model of human learning and memory. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 2, p. 170-215). Cambridge, MA: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1988). Explorations in parallel distributed processing: A handbook of models, programs, and exercises. Cambridge, MA: MIT Press.

McClelland, J. L., St. John, M. F., & Taraban, R. (1989). Sentence comprehension: A parallel distibuted processing approach. Language and Cognitive Processes, 4, 287-335.

Miikkulainen, R. (1993). Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory. Cambridge, MA: MIT Press.

Miikkulainen, R., & Dyer, M. G. (1991). Natural language processing with modular PDP networks and distributed lexicon. Cognitive Science, 15, 343-399.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. Psychological Review, 92, 289-316.

Pauen, S. (2002a). Evidence for knowledge-based category discrimination in infancy. Child Development, 73(4), 1016.

Pauen, S. (2002b). The global-to-basic shift in infants' categorical thinking: First evidence from a longitudinal study. International Journal of Behavioural Development, 26(6), 492-499.

Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. British Journal of Developmental Psychology, 10(3), 209-254.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), Semantic information processing (p. 227-270). Cambridge, MA: MIT Press.

Quine, W. V. O. (1960). Word and object. Cambridge, MA: MIT Press.

Quinn, P., & Eimas, P. (1997a). Perceptual organization and categorization in young infants. In C. Rovee-Collier & L. P. Lipsitt (Eds.), Advances in infancy research (Vol. 11). Norwood, NJ: Ablex.

Quinn, P., & Eimas, P. (1997b). A reexamination of the perceptual-to- conceptual shift in mental representations. Review of General Psychology, 1, 271-287.

Quinn, P., Eimas, P., & Rosenkrantz, S. (1991). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. Perception, 22, 463-475.

Quinn, P., & Eimas, P. D. (2000). The emergence of category representations during infancy: Are separate perceptual and conceptual processes really required? Journal of Cognition and Development, 1, 55-61.

Quinn, P. C. (2002). Early categorization: A new synthesis. In U. Goswami (Ed.), Blackwell handbook of childhood cognitive development (p. 84-101). Blackwell Publishers.

Quinn, P. C., & Johnson, M. H. (1997). The emergence of perceptual category representations in young infants: A connectionist analysis. Journal of Experimental Child Psychology, 66, 236-263.

Rakison, D. (in press). Parts, categorization, and the animate-inanimate distinction in infancy. In L. M. Oakes & D. H. Rakison (Eds.), Early concept and category development: Making sense of the blooming, buzzing confusion. New York: Oxford University Press.

Rakison, D., & Butterworth, B. (1998a). Infant attention to object structure in early categorization. Developmental Psychology, 34(6), 1310-1325.

Rakison, D., & Butterworth, B. (1998b). Infants' use of parts in early categorization. Developmental Psychology, 34(1), 49-62.

Rakison, D., & Cohen, L. B. (1999). Infants use of functional parts in basic-like categorization. Developmental Science, 2, 423-432.

Rakison, D., & Poulin-Dubois, D. (2001). The developmental origin of the animate-inanimate distinction. Psychological Bulletin, 127, 209-228.

Rogers, T. T., Lambon-Ralph, M., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (in press). The structure and deterioration of semantic memory: a computational and neuropsychological investigation.

Rogers, T. T., & McClelland, J. L. (in press). A parallel distributed processing approach to semantic cognition: Applications to conceptual development. In L. Gershkoff-Stowe & D. Rakison (Eds.), Building object categories in developmental time. Mahwah, NJ: Erlbaum.

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), An introduction to neural and electronic networks (p. 405-420). San Diego, CA: Academic Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(9), 533-536.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, & the PDP research group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 2, p. 7-57). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience (p. 3-30). Cambridge, MA: MIT Press.

Schyns, P. G. (1991). A modular neural network model of concept acquisition. Cognitive Science, 15, 461-508.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. Artificial Intelligence, 46, 217-257.

Younger, B., & Fearing, D. (2000). A global-to-basic trend in early categorization: Evidence from a dual-category habituation task. Infancy, 1, 47-58.

Zeki, S. M. (1978). Functional specialization in the visual cortex of the rhesus monkey. Nature, 274, 423.

Footnotes

[1]Scholars of philosophy and other students of Quine (1960) might observe that it is not a trivial matter for the child to determine what is being named when a name is given, and the argument can be extended to noting that it isn't at all clear what particular aspects of a situation are the ones that support the appropriate predictions of outcomes that might be expected to arise from it. Attentional forgrounding and characteristics of perceptual systems can influence these processes. Elsewhere we have argued as well that gradual learning in connectionist networks can sort out ambiguities that arise in individual cases (St. John & McClelland, 1990). For example, a naive listener may not know what aspects of an event are picked out by the words "cat" and "dog" in a sentence like "The dog is chasing the cat"; but over many other sentence-event pairs (e.g., events described with sentences "The dog is chewing the bone," "The cat is drinking the milk," etc.) some of these ambiguities will naturally sort themselves out.

[2]As a side point it should perhaps be noted that it is easy to manipulate salience explicitly in connectionist networks—for example, by introducing a scalar "salience" parameter that determines the strength with which each property can drive learning in the network. Rather than do this, our model might be seen instead as relying on the extreme approximation of a binary choice of salience values — salience of 1 for those properties included in the training set, salience of 0 for all others.

Figure Captions

Figure 1. A connectionist model of semantic memory adapted from Rumelhart and Todd (1993). The entire set of units used in the network is shown. Input units are shown on the left, and activation propagates from the left to the right. Where connections are indicated, every unit in the pool on the left is connected to every unit in the pool to the right. Each unit in the Item layer corresponds to an individual item in the environment. Each unit in the Relation layer represents contextual constraints on the kind of information to be retrieved. Thus, the input pair robin can corresponds to a situation in which the network is shown a picture of a robin, and asked what it can do. The network is trained to turn on all those units that represent correct completions of the input query. In the example shown, the correct units to activate are grow, move, and fly. Based on the network depicted in Rumelhart and Todd (1993), Figure 1.9, page 15. Permission pending.

Figure 2. Learned internal representations of eight items at three points during learning, using the network shown in Figure 1. The height of each vertical bar indicates the degree of activation for one of the eight units in the network's Representation layer, in response to the activation of a single Item unit in the model's input. Early in learning (50 Epochs), the pattern of activation across these units is similar for all eight objects. After 100 epochs of training, the plants are still similar to one another, but are distinct from the animals. By 150 epochs, further differentiation into trees and flowers is evident. Similar results were previously reported in McClelland, McNaughton and O'Reilly (1995).

Figure 3. Hierarchical cluster plot of the learned internal representations at three points during training. The cluster analysis makes the similarity relations shown in Figure 2 explicit. The clustering algorithm recursively links a pattern or a previously-linked group of patterns to another pattern or previously-formed group. The process begins with the pair that is most similar, whose elements are then replaced by the resulting group. These steps are repeated until all items have been joined in a single superordinate group. Similarity is measured by the Euclidean distance metric. The results show that the network is first sensitive to broad semantic distinctions, and only gradually picks up on more specific ones. Similar results have previously been reported by McClelland et al. (1995) and Quinn and Johnson (1997).

Figure 4. Trajectory of learned internal representations during learning. The Euclidean distance matrix for all item representations was calculated at ten different points throughout training. A multidimensional scaling

was performed on these data to find corresponding points in a two dimensional space that preserve, as closely as possible, the pairwise distances among representations across training. Thus, the proximity of two points in the figure approximates the actual Euclidean distance between the network's internal representations of the corresponding objects at a particular point in training. The lines indicate the path traversed by a particular item representation over the course of development.

Figure 5. Bottom: Mean distance between plant and animal, bird and fish, and canary and robin internal representations throughout training. Middle: average magnitude of the error signal propagating back from properties that reliably discriminate plants from animals, birds from fish, or the canary from the robin, at different points throughout training when the model is presented with the canary as input. Top: Activation of a property shared by animals (can move), birds can fly or unique to the canary (can sing), when the model is presented with the input canary can at different points throughout training.

Figure 6. Base architecture used for for our simulation of conceptual differentiation in infancy.

Figure 7. Matrix indicating the similarities among the different items in their attribute structures. The shading indicates the degree to which two items have similar attributes, with dark colors signifying items with higher degrees of similarity. The measure of similarity is obtained by concatinating the output patterns for the is, can, and has contexts into a single vector for each item, with 1's for attributes that are present and 0's for attributes that are absent, then computing the normalized dot-product among the item representations.

Figure 8. Hierarchical cluster plot of the similarities expressed by the overlap of input features used in the distributed-inputs implmentation.

Figure 9. Hierarchical cluster plot of the internal representations acquired by the model at three different points in learning. The distance matrices were generated by calculating the mean pairwise Euclidean distances among representations across five different network runs in each implementation.

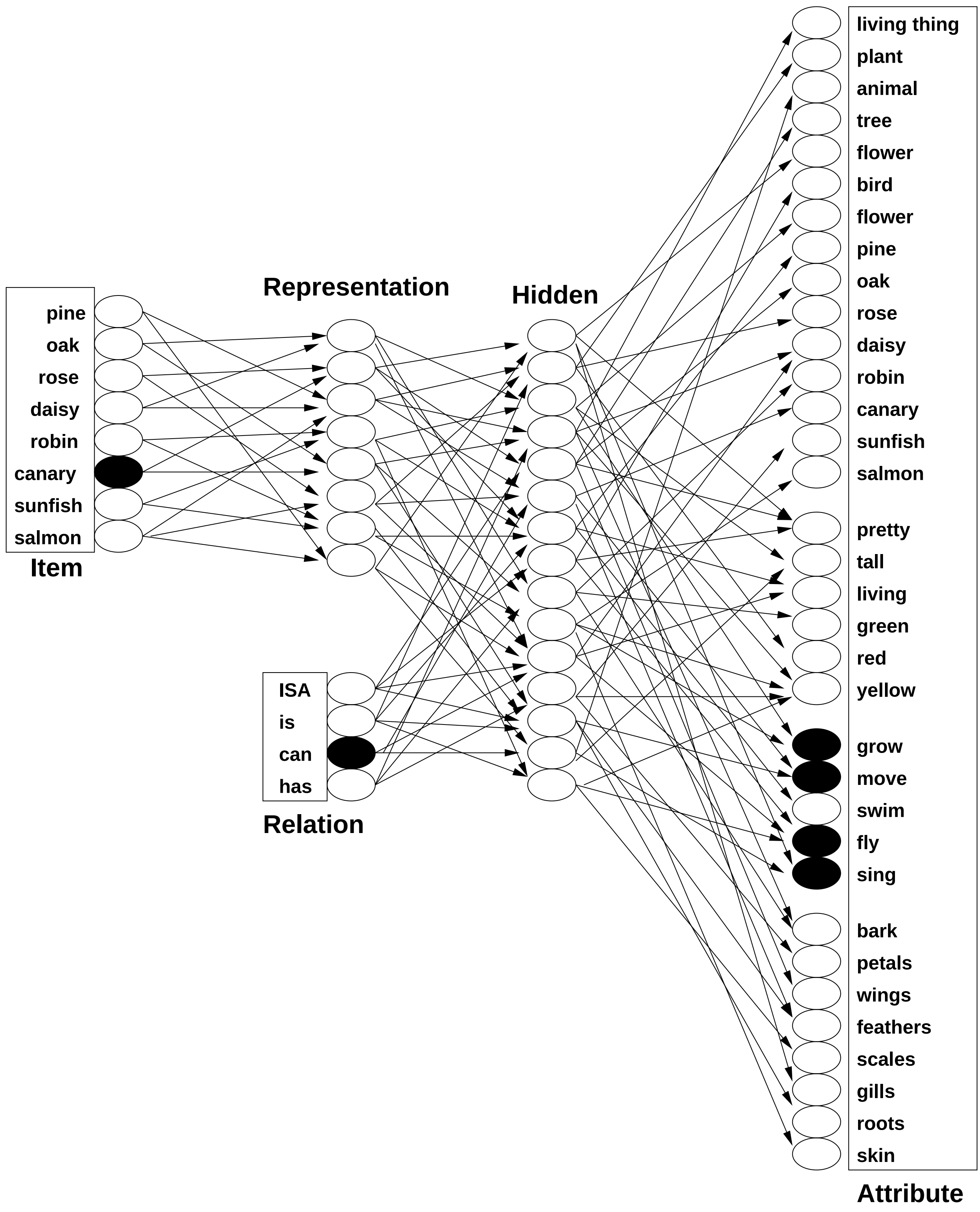
Figure 10. Attributes of 20 novel items used to simulate Mandler and McDonough's (1993) infant-preference experiment. Associated with each category are four perceptually similar exemplars which constitute habituation items, one perceptually dissimilar exemplar employed as a test item, and two perceptually similar out-of-category test items: one from the contrasting category in the same domains, and one in the opposite

domain.

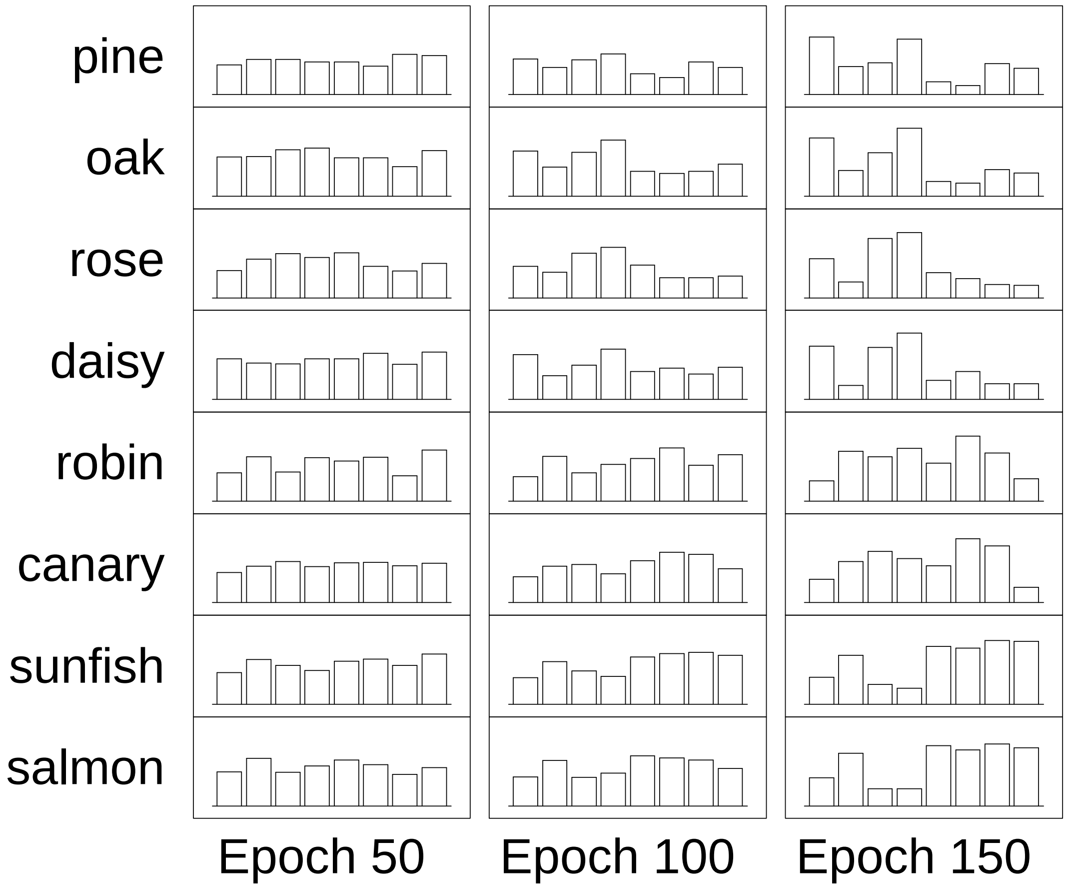
Figure 11. Hierarchical cluster plot of the perceptual similarities for some of the novel habituation and test items used in the simulation. In each case, the fifth category item is less similar to its category coordinates than is an out-of-category test item.

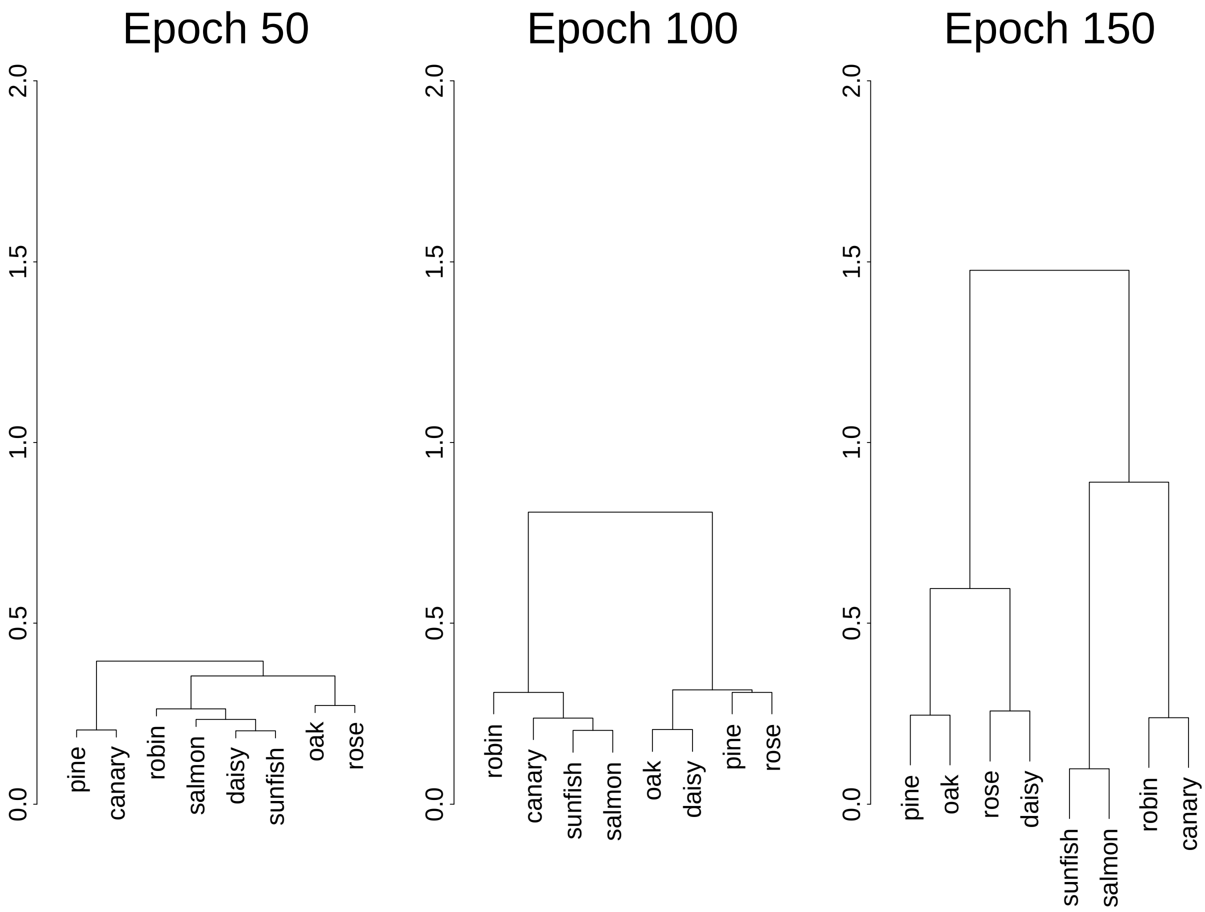
Figure 12. Hierarchical cluster plot of the representations generated by the distributed-inputs model, for novel habituation and test items from the “bird” category at three different points during training.

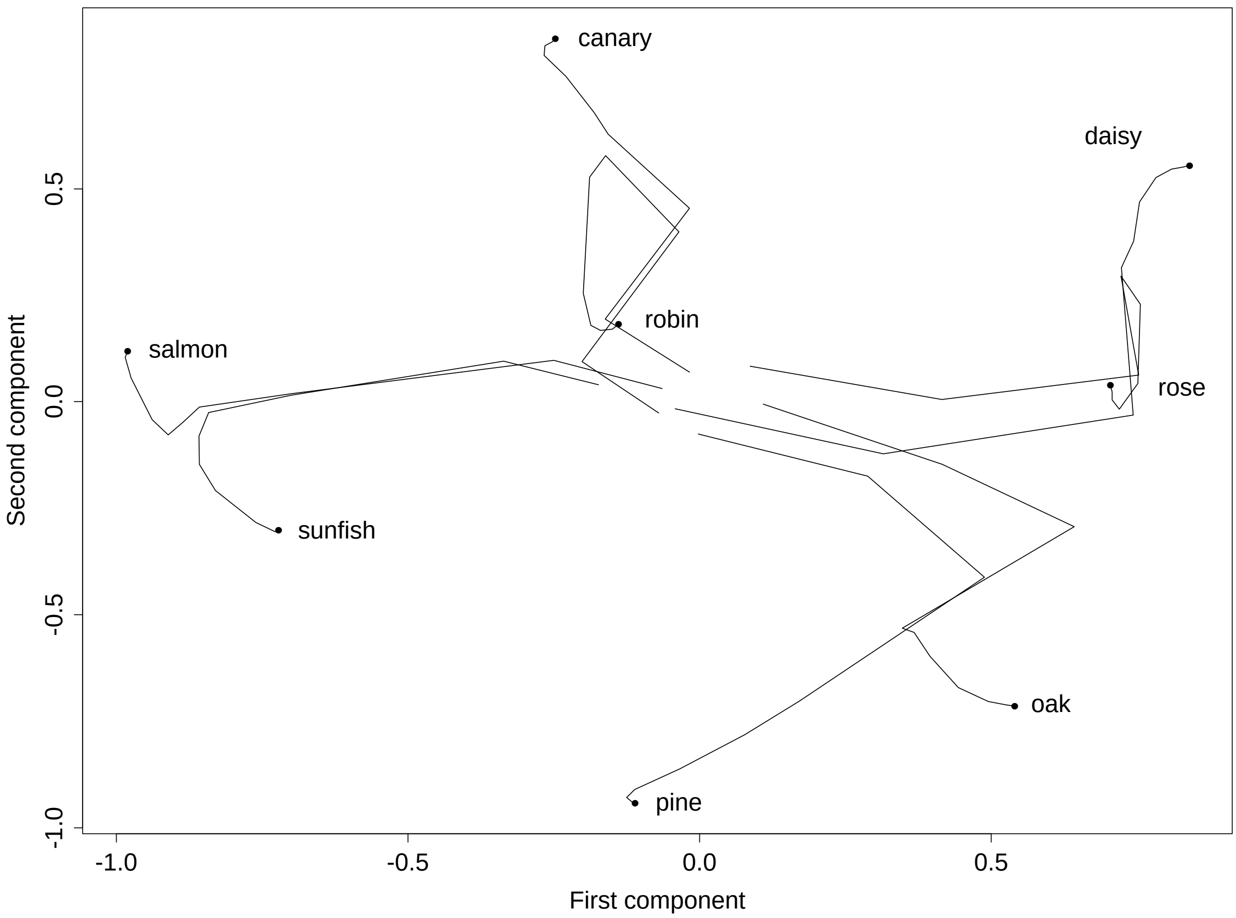
Figure 13. Average likelihood of “preferring” the semantically unrelated test item througout early epochs of learning, for “global” and “basic” semantic contrasts, in a model analog of the object examination task.

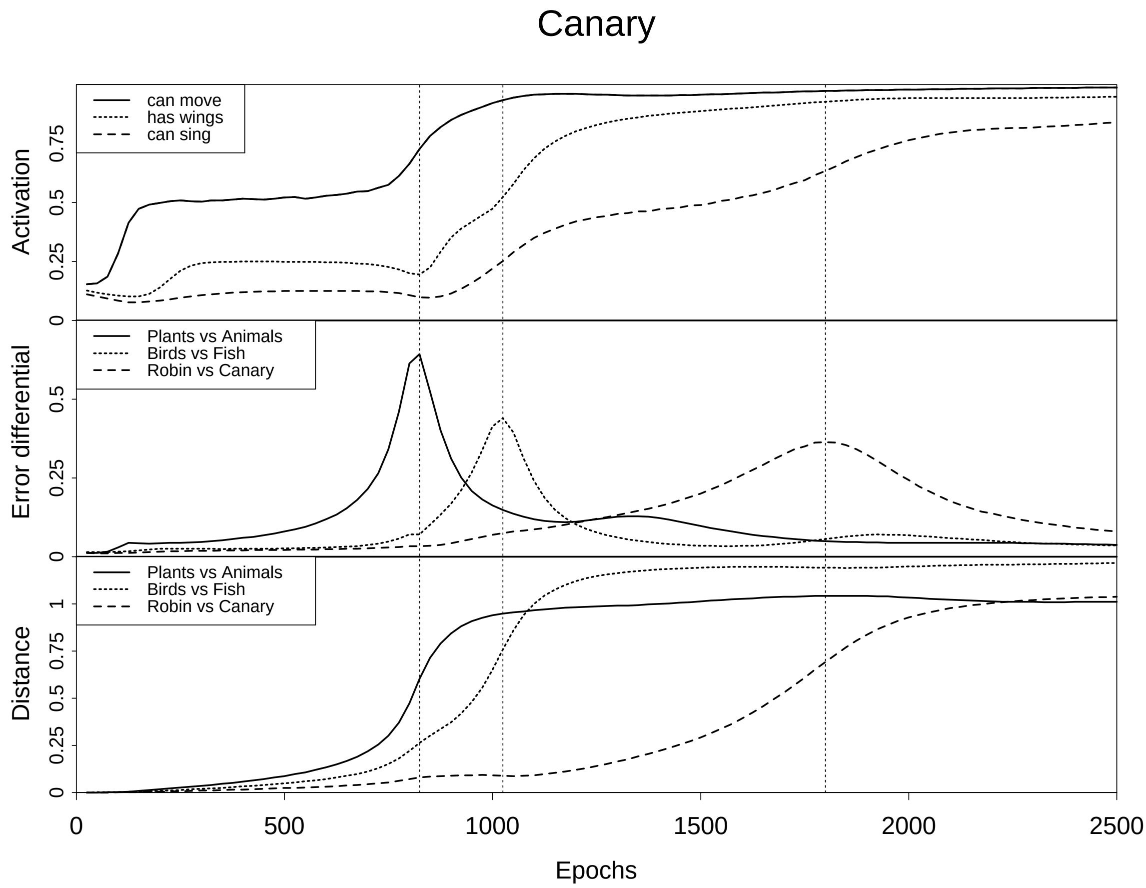


Representation
Hidden
Item
pine
oak
rose
daisy
robin
canary
sunfish
salmon
Relation
ISA
is
can
has
Attribute
living thing
plant
animal
tree
flower
bird
flower
pine
oak
rose
daisy
robin
canary
sunfish
salmon
pretty
tall
living
green
red
yellow
grow
move
swim
fly
sing
bark
petals
wings
feathers
scales
gills
roots
skin

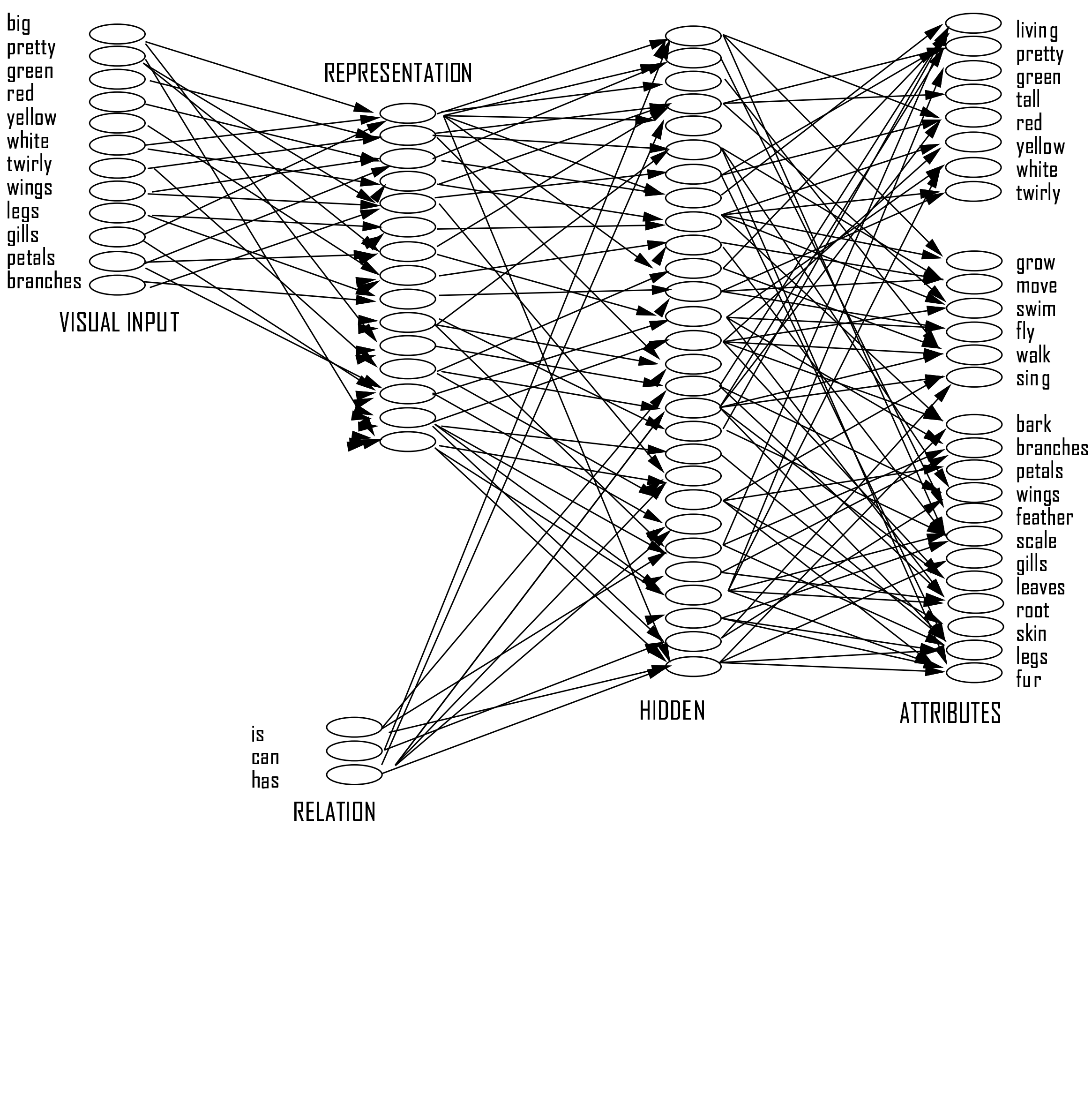

pine
oak
rose
daisy
robin
canary
sunfish
salmon
Epoch 50
Epoch 100
Epoch 150

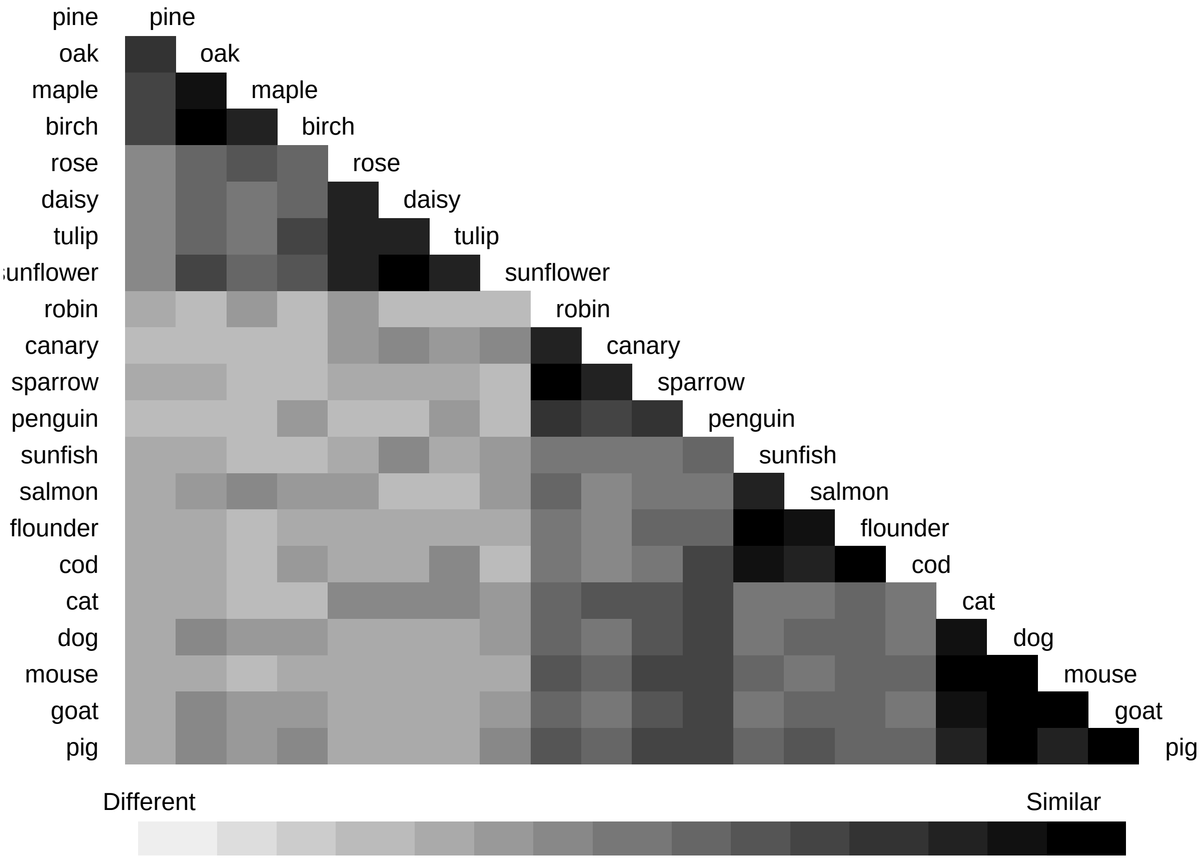

Epoch 50
Epoch 100
Epoch 150
0.0
0.5
1.0
1.5
2.0
pine
canary
robin
salmon
daisy
sunfish
oak
rose
robin
canary
sunfish
salmon
oak
daisy
pine
rose
pine
oak
rose
daisy
sunfish
salmon
robin
canary

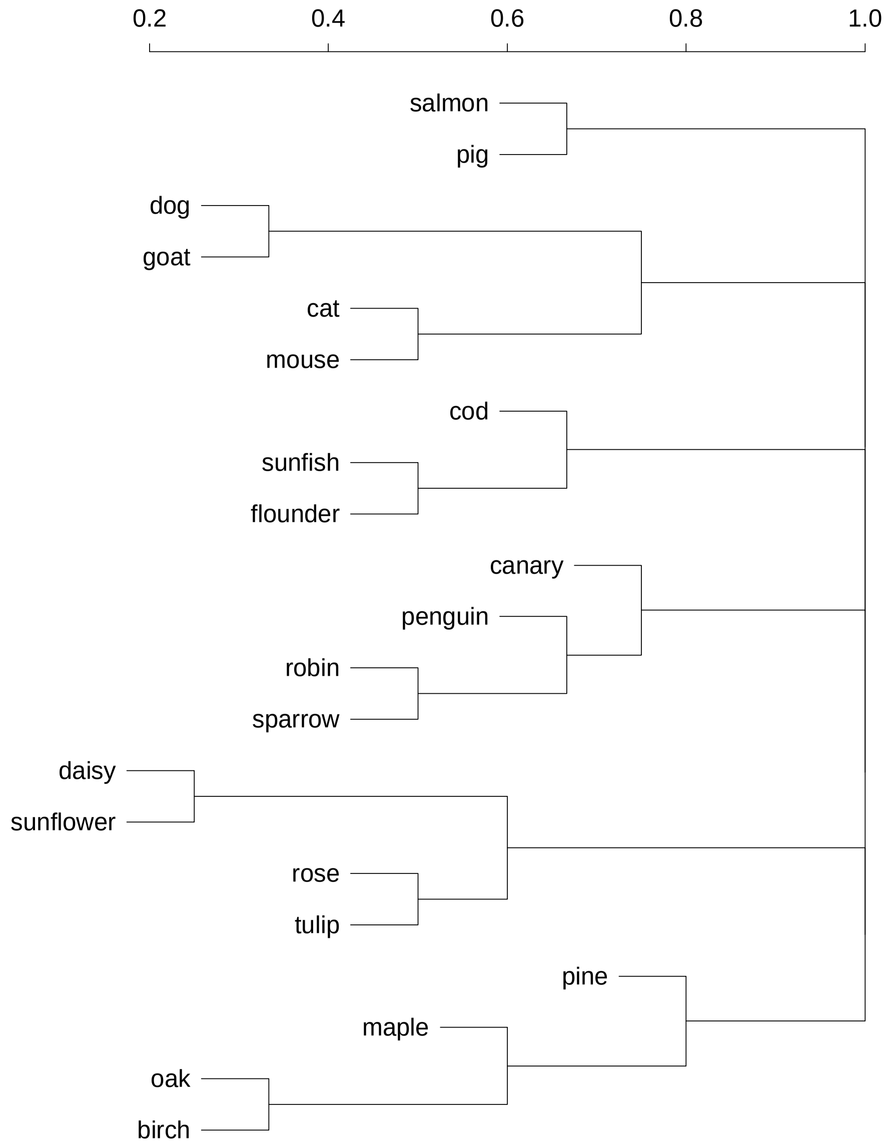

canary
daisy
robin
salmon
rose
sunfish
oak
pine
0.5
0.0
-0.5
-1.0
Second component
-1.0
-0.5
0.0
0.5
First component

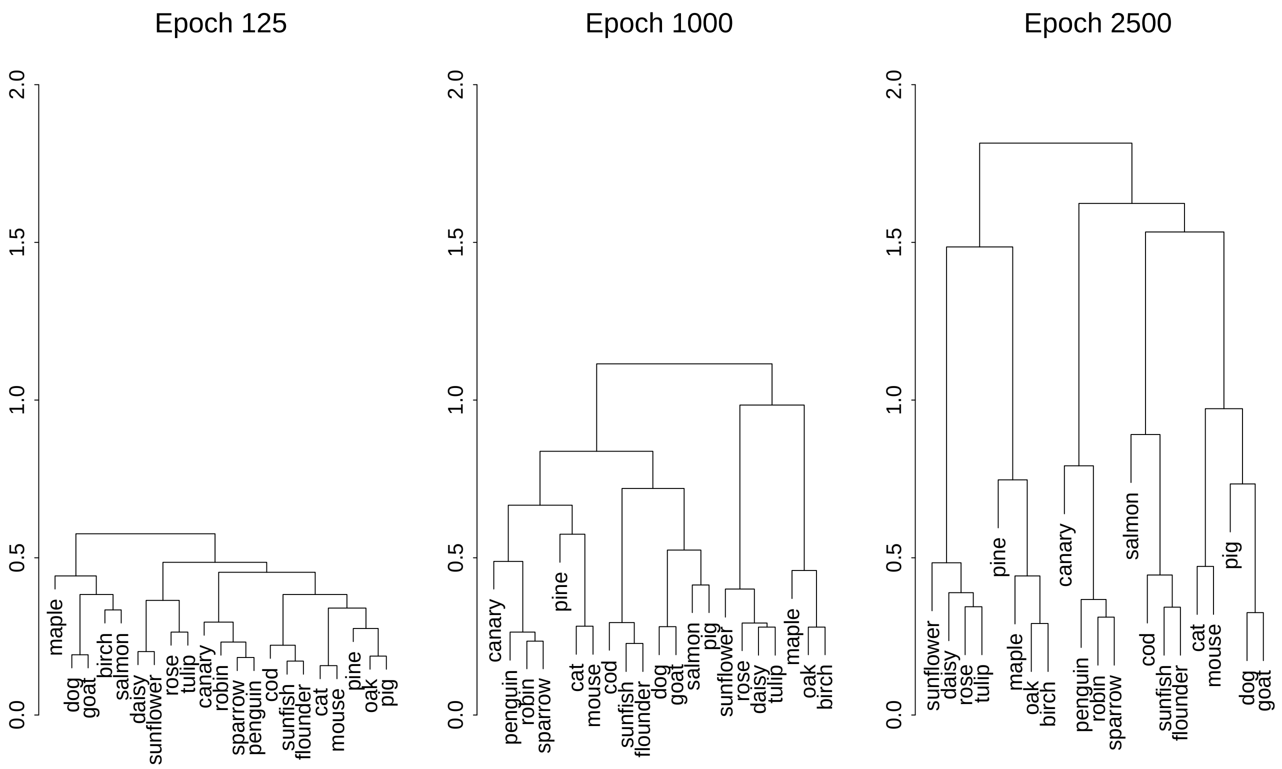


Canary
can move
has wings
can sing
Activation
0
0.25
0.5
0.75
Plants vs Animals
Birds vs Fish
Robin vs Canary
Error differential
0
0.25
0.5
Plants vs Animals
Birds vs Fish
Robin vs Canary
Distance
0
0.25
0.5
0.75
1
0
500
1000
1500
2000
2500
Epochs


big
pretty
green
red
yellow
white
twirly
wings
legs
gills
petals
branches
VISUAL INPUT
REPRESENTATION
is
can
has
RELATION
HIDDEN
living
pretty
green
tall
red
yellow
white
twirly
grow
move
swim
fly
walk
sing
bark
branches
petals
wings
feather
scale
gills
leaves
root
skin
legs
fur
ATTRIBUTES

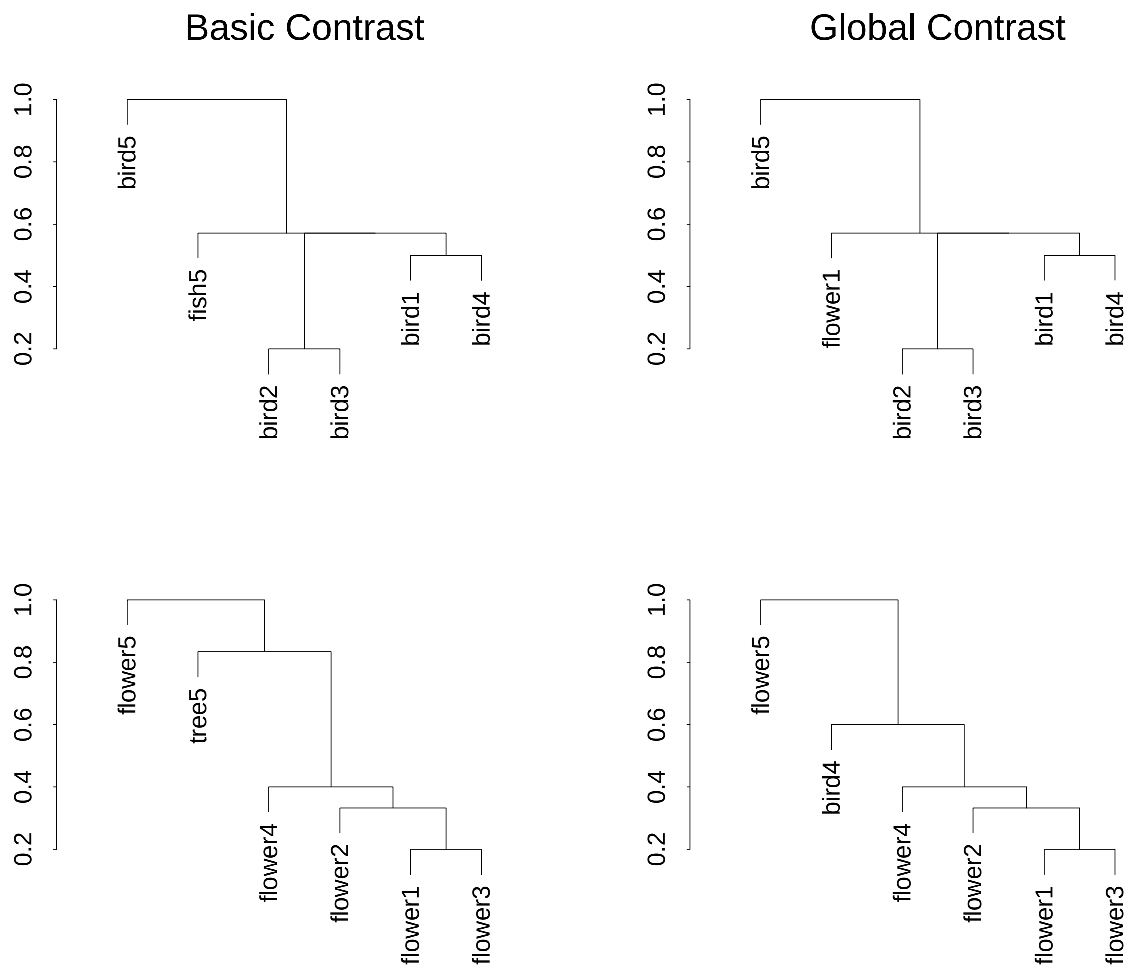

pine
oak
maple
birch
rose
daisy
tulip
sunflower
robin
canary
sparrow
penguin
sunfish
salmon
flounder
cod
cat
dog
mouse
goat
pig
Different
Similar

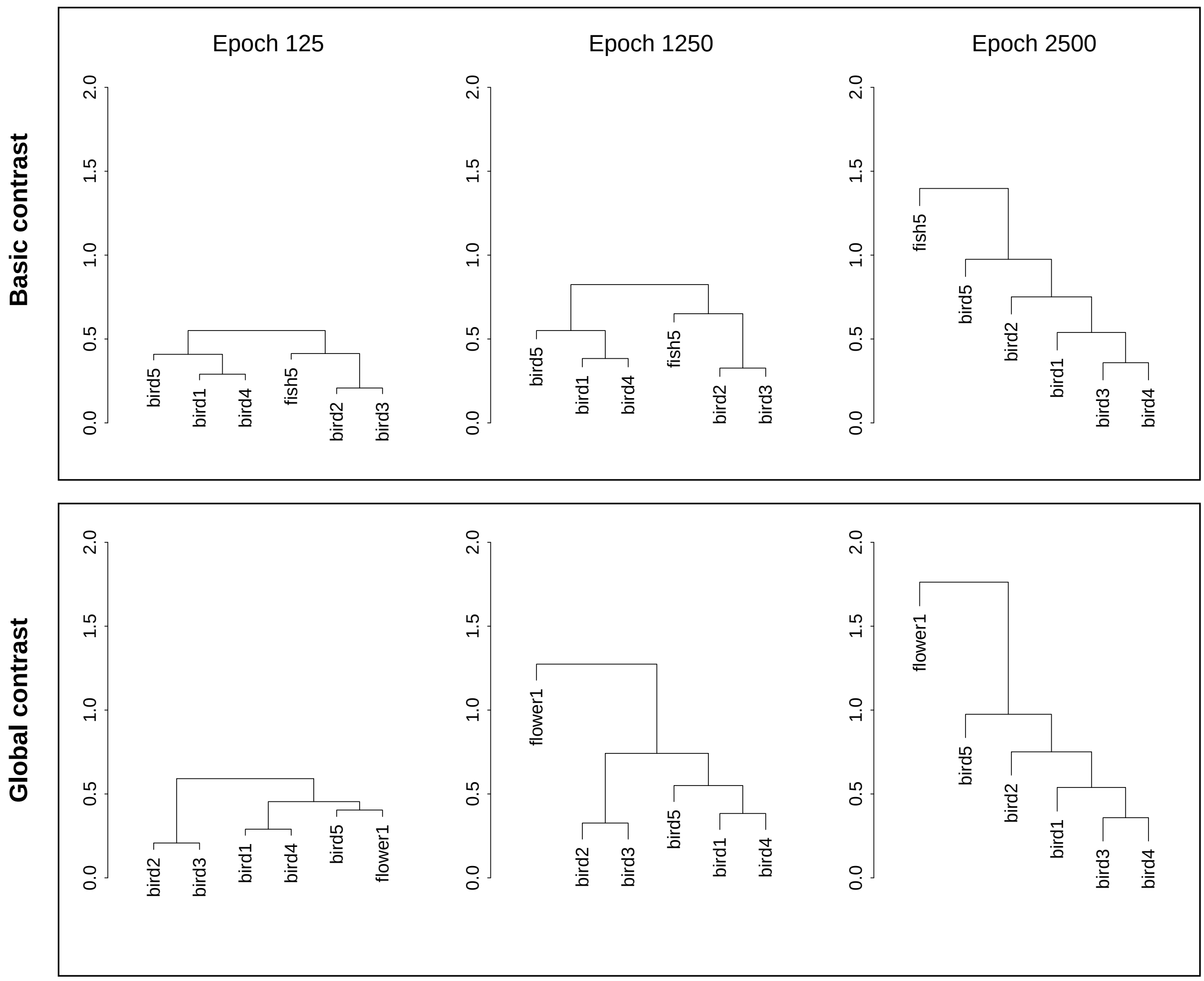

0.2
0.4
0.6
0.8
1.0
salmon
pig
dog
goat
cat
mouse
cod
sunfish
flounder
canary
penguin
robin
sparrow
daisy
sunflower
rose
tulip
pine
maple
oak
birch

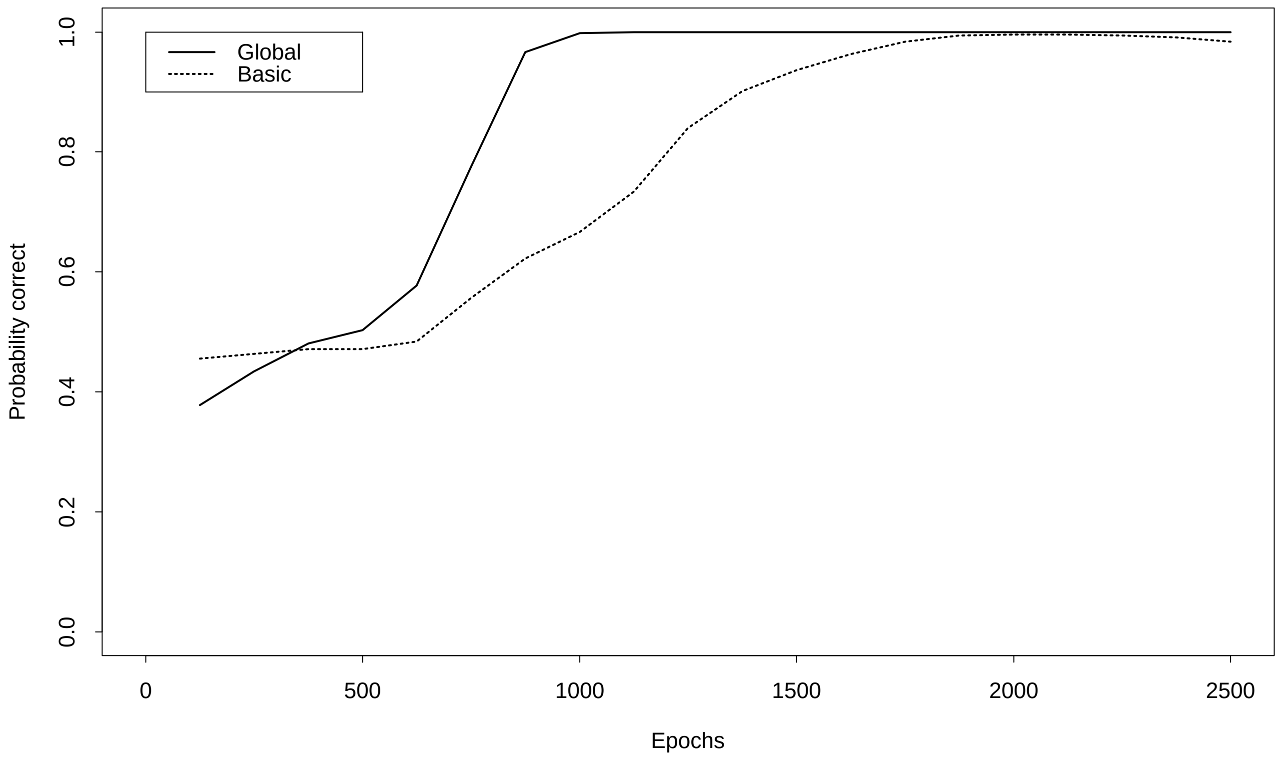


Epoch 125
Epoch 1000
Epoch 2500
0.0
0.5
1.0
1.5
2.0
maple
dog
goat
birch
salmon
daisy
sunflower
rose
tulip
canary
robin
sparrow
penguin
cod
sunfish
flounder
cat
mouse
pine
oak
pig
canary
penguin
robin
sparrow
pine
cat
mouse
cod
sunfish
flounder
dog
goat
salmon
pig
sunflower
rose
daisy
tulip
maple
oak
birch
sunflower
daisy
rose
tulip
pine
maple
oak
birch
canary
penguin
robin
sparrow
salmon
cod
sunfish
flounder
cat
mouse
pig
dog
goat

	pretty	big	green	red	yellow	white	twirly	branches	petals	wings	gills	fur
tree1	0	1	1	1	1	0	0	1	0	0	0	0
tree2	0	1	1	0	1	1	0	1	0	0	0	0
tree3	0	1	1	0	1	0	0	1	0	0	0	0
tree4	0	1	1	0	0	0	0	1	0	0	0	0
tree5	0	0	0	1	0	1	1	1	0	0	0	0
flower1	1	0	0	1	0	1	1	0	1	0	0	0
flower2	1	0	0	0	1	1	1	0	1	0	0	0
flower3	1	0	0	0	0	1	1	0	1	0	0	0
flower4	1	0	0	0	0	1	0	0	1	0	0	0
flower5	0	1	1	0	1	0	0	0	1	0	0	0
bird1	1	0	1	1	0	0	1	0	0	1	0	0
bird2	1	0	0	1	1	1	0	0	0	1	0	0
bird3	1	0	0	1	0	1	0	0	0	1	0	0
bird4	1	0	0	0	0	1	1	0	0	1	0	0
bird5	0	1	1	0	1	0	0	0	0	1	0	0
fish1	0	1	1	0	0	1	0	0	0	0	1	0
fish2	0	1	1	1	1	0	0	0	0	0	1	0
fish3	0	1	1	0	1	0	1	0	0	0	1	0
fish4	0	1	1	0	1	0	0	0	0	0	1	0
fish5	1	0	0	1	0	1	1	0	0	0	1	0


Basic Contrast
Global Contrast
1.0
0.8
0.6
0.4
0.2
bird5
fish5
bird2
bird3
bird1
bird4
flower1
flower5
tree5
flower4
flower2
flower3
bird4


Basic contrast
Global contrast
Epoch 125
Epoch 1250
Epoch 2500
0.0
0.5
1.0
1.5
2.0
bird5
bird1
bird4
fish5
bird2
bird3
flower1


Global
Basic
Probability correct
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Epochs