Cell PRESS

Approaches to cognitive modeling

Letting structure emerge: connectionist and dynamical systems approaches to cognition

James L. McClelland[1], Matthew M. Botvinick[2], David C. Noelle[3], David C. Plaut[4], Timothy T. Rogers[5], Mark S. Seidenberg[5] and Linda B. Smith[6]

[1] Department of Psychology, Stanford University, Building 420, 450 Serra Mall, Stanford, CA 94305, USA
[2] Department of Psychology and Princeton Neuroscience Institute, Princeton University, Green Hall, Princeton, NJ 08504, USA
[3] School of Engineering and School of Social Sciences, Humanities, and Arts, Merced, 5200 North Lake Road, University of California, Merced, CA 95343, USA
[4] Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[5] Department of Psychology, University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706, USA
[6] Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th Street, Bloomington, IN 47405, USA

Connectionist and dynamical systems approaches explain human thought, language and behavior in terms of the emergent consequences of a large number of simple noncognitive processes. We view the entities that serve as the basis for structured probabilistic approaches as abstractions that are occasionally useful but often misleading: they have no real basis in the actual processes that give rise to linguistic and cognitive abilities or to the development of these abilities. Although structured probabilistic approaches can be useful in determining what would be optimal under certain assumptions, we propose that connectionist, dynamical systems, and related approaches, which focus on explaining the mechanisms that give rise to cognition, will be essential in achieving a full understanding of cognition and development.

Emergence of structure in cognition

Emergence is ubiquitous in nature: consider the complex structure of an anthill. It can have a complex architecture, with a complex network of passageways leading from deep underground to 7.5 m into the sky. One might suppose that ants possess a blueprint for creating such structures, but something far simpler is in play [1]. Ants are sensitive to certain gasses within their nests; when these gasses build up they move grains of dirt to the outside. This activity lets the gasses escape and has the byproduct of creating the elaborate structure of the nest.

Likewise, human thoughts and utterances have a rich and complex structure that, in our view, is also the emergent consequence of the interplay of much simpler processes. The emergentist view contrasts with the approach advocated in the companion article [2], in which cognizing agents are viewed as optimal inferencing machines, coming to cognitive tasks with a structured hypothesis space and a prior probability distribution over hypotheses. Observations provide a means of evaluating the hypotheses and selecting the one that has the highest posterior probability. Work within the structured probabilistic framework is often thought to address an abstract level of analysis akin to Marr's computational level [3], with consideration of the actual cognitive

Glossary

Connectionism: An approach to modeling cognition based on the idea that the knowledge underlying cognitive activity is stored in the connections among neurons. In connectionist models, knowledge is acquired by using an experience-driven connection adjustment rule to alter the strengths of connections among neuron-like processing units.
Dynamical field theory: Originally formulated as a theory of movement preparation, in which movement parameters are represented by distributions of activation defined over metric spaces, the theory has recently been extended to address cognitive function. Dynamical fields are formalizations of how neural populations represent the continuous dimensions that characterize perceptual features, movements and cognitive decisions, and dynamical field theory specifies how activity in such neural populations evolves over time.
Dynamical system: A mathematical formalization that describes the time evolution of physical and cognitive states. Examples include the mathematical models that describe the swinging of a clock pendulum, the flow of water in a pipe, the movement of the limbs of a walking organism, and the drift that occurs in working memory towards or away from special points in the state space.
Emergentist approaches: Approaches to modeling cognition based on the idea that the structure seen in overt behavior and the patterns of change observed in behavior reflect the operation of subcognitive processes such as propagation of activation and inhibition among neurons and adjustment of strengths of connections between them. In contrast to emergentist approaches, symbolic approaches, including structured probabilistic models, model cognition directly at the level of manipulation of symbols and symbolic structures such as propositions and rules.
Semantic cognition: A cognitive domain encompassing knowledge of the properties of objects and their relationships to other objects, as well as the acquisition of such knowledge and its use in guiding inference.
Structured probabilistic models: Models that specify that cognitive activity involves the use of probabilistic information to select among and specify the parameters of particular structural forms that specify relationships among items represented by discrete symbols.
Universal grammar: A hypothetical construct that arose in the context of generative grammar. A universal grammar, if one existed, would be an idealized structured representation that captures properties shared by all natural languages.

E-mail addresses: mcclelland@stanford.edu

1364-6613/$ – see front matter doi:10.1016/j.tics.2010.06.002 Trends in Cognitive Sciences 14 (2010) 348–356

Box 1. Parallel pitfalls of computational-level and competence approaches

Structured probabilistic inference models include the following elements:

- Formulation of any given problem as one of probabilistic inference.
- Commitment to selecting the correct knowledge structure over which probabilities can be assessed and updated.
- Abstraction from details of behavior and brain because the theory is usually pitched at Marr's computational level.

A broader perspective on this approach is provided by looking at its closely-related precursor, Chomsky's competence-based approach to linguistics [4], whose foundational assumptions included the following:

- Formulation of the goal of the field as characterizing a language user's knowledge.
- Commitment to selecting the correct grammar as the representation that explains such facts.
- Abstraction from details of behavior and brain because the theory is pitched at the competence level.

In both cases, the goal is an abstract characterization; linkage to performance is a promissory note, seldom redeemed in practice.

Thus, structured probabilistic models of cognition can be understood as competence theories. As such they inherit problems that have become apparent with this approach, including:

- The problem formulation is not neutral. If learners are not trying to 'select the correct grammar' or 'the correct structure' for a domain, and approach the problem as if they were, this would be misleading.
- The commitment to a form of knowledge representation is not neutral. Commitments to particular choices can lead researchers into a blind alley. Commitment to grammar formalisms radically constrains how other issues are addressed. Acquisition becomes the problem of converging on a grammar, performance the question of how grammar is used and neurolinguistics the study of how grammar is represented in the brain. The role of grammatical theory has greatly diminished over the years because of the research program's lack of progress.
- Treating levels of analysis as independent is counterproductive. It might be difficult or impossible to relate the high-level computational/competence theory back to facts about behavior and the brain. Conversely, considering implementation/performance issues can lead to a different high-level formulation of a problem.
- The levels of description and competence/performance approaches also introduce an uncomfortable extra degree of freedom with respect to data. Facts that are consistent with the theory are embraced whereas facts that conflict with the theory are relegated to as yet undeveloped 'algorithmic-implementational' or 'performance' theories.

processes being deferred until the computational-level theory is fully worked out.

The danger, of course, is that if the high-level description is wrong – that is, if the behaving child or adult were not actually engaged in the formulation and selection of hypotheses – then focusing on these constructs would be misleading. It could give rise to an enterprise, similar to Chomsky's competence theory of universal grammar [4], in which researchers focus on searching for entities that might exist only as descriptive abstractions, while ignoring those factors that actually shape behavior (Box 1).

Explanations of behavior that ignore mechanism and implementation are likely to fall short. For example, a recent study [5] has found that people can exploit a causal framing scenario to make normatively correct, explicit inferences in a contingency learning task if they are given

Box 2. The units problem in language and cognition

Language is usually characterized in terms of discrete units such as phonemes, morphemes and sentences. Such units are compatible with probabilistic inference models that employ structured representations. For example, recognizing a speech sound could be construed as a Bayesian inference problem in which the hypotheses are alternative phonemes and the task is to pick the one that is most probable given the input [35]. The utility of this approach depends in part on the validity of the units as descriptions of linguistic structure. Herein lies a problem.

All of these units can be intuitively motivated using apparently clear cases: phonemes are illustrated by minimal pairs such as PEN and TEN, and morphemes are minimal units of meaning as in FARM–FARMER. Such units provide useful terminology for describing and comparing. However, it would be a mistake to take them as the units involved in acquiring and using language.

In actual spoken language, units such as phonemes and syllables are matters of degree. There is almost no 't' in 'softly', but more of one in 'swiftly' [36]; words such as 'memory' have more than two syllables but less than three [37]. Morphology presents a similar problem. There are cases in which the meaning of a complex word seems to be compositional (prefabricate), others where there is no compositionality at all (corner), and still others (predict, prefer) in which the parts seem to contribute to, but do not fully determine, the meaning of the whole [38]. Data suggest that people are sensitive to the gradations, in that intermediate cases produce intermediate morphological priming effects [39], indicating that morphological status is a matter of degree. For years, syntactic theory treated sentences as grammatical or ungrammatical. However, the borderline cases are legion [40]. In light of such observations, many linguists have turned to formalisms that admit degrees of well formedness [41,42]. However, these systems still generally require commitments to a set of units over which degrees of well formedness can be computed. Similar issues arise in all efforts to create a taxonomy of concepts or meanings for words.

In connectionist models, there is no fixed vocabulary of representational units. The internal representations are graded patterns with varying degrees of distinctness, compositionality and context sensitivity [43–45]. These characteristics make connectionist models different from a mere 'implementation' of an idealized linguistic theory.

ample time to make explicit predictions. However, when the same contingencies govern events to which participants must respond very quickly, they seem to learn according to a process akin to simple connection weight adjustment. Thus, different mechanisms seem to underlie learning of the very same probabilistic contingencies in the explicit prediction versus quick response variants of the task, yet the statistical structure of the two tasks, and thus the computational-level analysis of what would be optimal in the two situations, is the same.

To be clear, the disagreement between emergentist approaches and structured probabilistic approaches is not about the relevance of probability in characterizing human behavior: both approaches share an emphasis on statistical regularities in the learning environment and on variability in human performance. Indeed, emergentist models often optimize their probabilistic behavior by learning to match probabilistic outputs to the statistical structure of the experiences on which they are trained [6,7]. The disagreement is also not about advocating a purely bottom-up versus top-down research strategy because it is our view that science is best served by pursuing integrated accounts that span multiple levels of analysis simultaneously. Rather, the dispute between the two approaches concerns the utility of treating cognition as if its goal and outcome is

Box 3. Examples of emergent phenomena in language, development and cognition

Language

Past-tense inflection and single word reading: Systematic linguistic knowledge (e.g. the past tense of BAKE is BAKED) is often attributed to the operation of explicit rules, with violations (TAKE/TOOK) relegated to separate, item-specific storage [46]. Connectionist approaches in domains including past-tense inflection [47,48] and single word reading [44,49] have emphasized instead that linguistic structure is graded rather than all-or-none, and that the relevant empirical phenomena are better captured by an integrated system in which all types of items are represented and processed.

Sentence processing: Classical approaches assume an innate module imbued with Universal Grammar as the basis for acquisition of syntactic knowledge. However, Elman [21,50] addressed the acquisition of syntax in a simple and generic connectionist model call the Simple Recurrent Network (SRN) (Figure I). Work by Elman and others has shown how SRNs can assign representations to words that capture their syntactic and semantic roles in sentences and respect subtle regularities including long-distance dependencies without explicit syntactic rules [51]. Related models learn to comprehend sentences and stories ([52,53]; see also Rohde, D.L.T., 2002, unpublished PhD thesis, School of Computer Science, Carnegie Mellon University).

Development

Stage transitions: It has been common to characterize development as occurring through a series of discrete stages. However, there are many signs that stage transitions are graded rather than discrete [54,55]. Connectionist models address such transitions as consequences of nonlinearities in multilayer networks. Effects of connection-weight changes in such networks exhibit accelerations and plateaus capturing stage-like phenomena [56,57].

U-shaped developmental trajectories: Young babies held upright seem to walk, but this behavior ceases long before self-supported walking. Classical accounts explain the disappearance as reflecting development of top-down inhibition [58]. More recent research shows that the disappearance reflects an increase in the mass of the child's legs as they develop [59]. The emergentist approach correctly predicts that walking can be evoked after its apparent disappearance with appropriate adjustments to counterbalance the effects of increased leg mass.

Cognitive processes

Semantic cognition: A connectionist model [60] accounted for apparent modular representation of living things versus artifacts as an emergent consequence of representation of visual and functional properties, and greater importance of functional properties for artifacts and of visual properties for living things (see also [20,27]).

Executive functions and short-term memory: The control of behavior by task and previous context is disrupted in individuals with brain lesions in a wide range of brain areas, even though such control has been ascribed to special modules in the frontal lobes [61]. Botvinick and Plaut [62] observed that when complex behaviors have been acquired by a generic SRN, diffuse damage leaves stereotyped action patterns intact but distrupts 'control' by task and context, indicating that such control could be an emergent function distributed over contributing brain areas. Their model also learns hierarchically structured tasks without explicitly representing hierarchical structure. Botvinick and Plaut [63] applied a similar model to a range of short-term memory phenomena that other approaches interpret as evidence for slots in short-term memory. In their model, the phenomena arise without explicit slots.

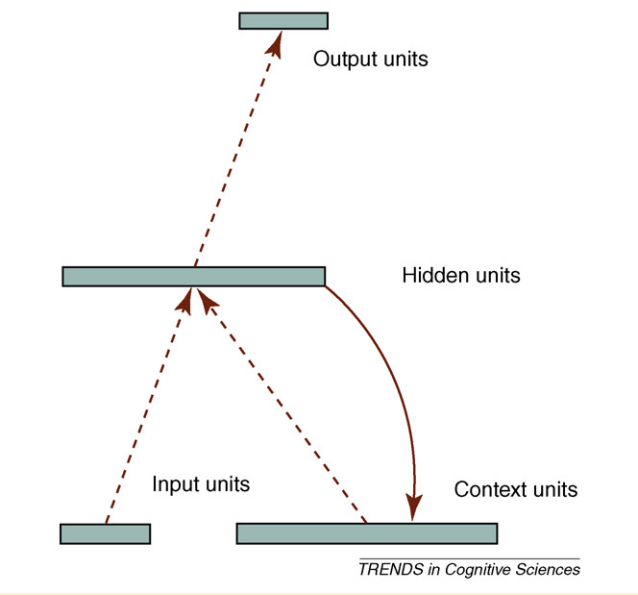

Figure I. Elman's simple recurrent network. Each rectangle represents a pool of simple processing units, and each dashed arrow represents a set of learnable connections from the units in one pool to the units in another. A stream of items is presented to the input layer of the network, one after another. For each item, the task is to predict the next item. The pattern on the hidden layer from processing the previous item is copied back to the context layer, thereby allowing context to influence the processing of the next incoming item. Reproduced, with permission, from Ref. [21].

the selection of one or the other structured statistical model, whether it be a probabilistic grammar, a mutation hierarchy, or a specific causal Bayes network [8–10]. From our perspective, the hypotheses, hypothesis spaces and data structures of the structured probabilistic approach are not the building blocks of an explanatory theory. Rather, they are sometimes helpful but often misleading approximate characterizations of the emergent consequences of the real underlying processes. Likewise, the entities over which these hypotheses are predicated – such as concepts, words, morphemes, syllables and phonemes – are themselves best understood as sometimes useful but sometimes misleading approximations (Box 2).

The remaining sections consider two very different cognitive domains that have been modeled as emergent phenomena using connectionist and dynamical systems approaches. In each case, we argue that it is unnecessary, and could even lead research astray, to characterize the situation in terms of structured probabilistic inference. In Box 3 we list examples of other linguistic, developmental and cognitive domains where the phenomena have been captured within emergentist approaches.

The A-not-B error: absence of a hypothesis or emergent consequence of the dynamics of motor behavior?

The A-not-B task was introduced by Piaget [11] to measure the development of the object concept: the belief that objects exist independent of one's own actions. In the canonical form of the task (Figure 1), after searching for an object at one location, then seeing it hidden at a new location, 8–10-month-old infants reach back to that first location, whereas older infants reach correctly to the new location. Although the A-not-B task has not been an explicit focus of research within the structured probabilistic framework, the situation is traditionally described in a way that is fully consistent with it: on this view, the phenomenon reflects the absence of (or perhaps a low prior probability for) the hypothesis that the object exists

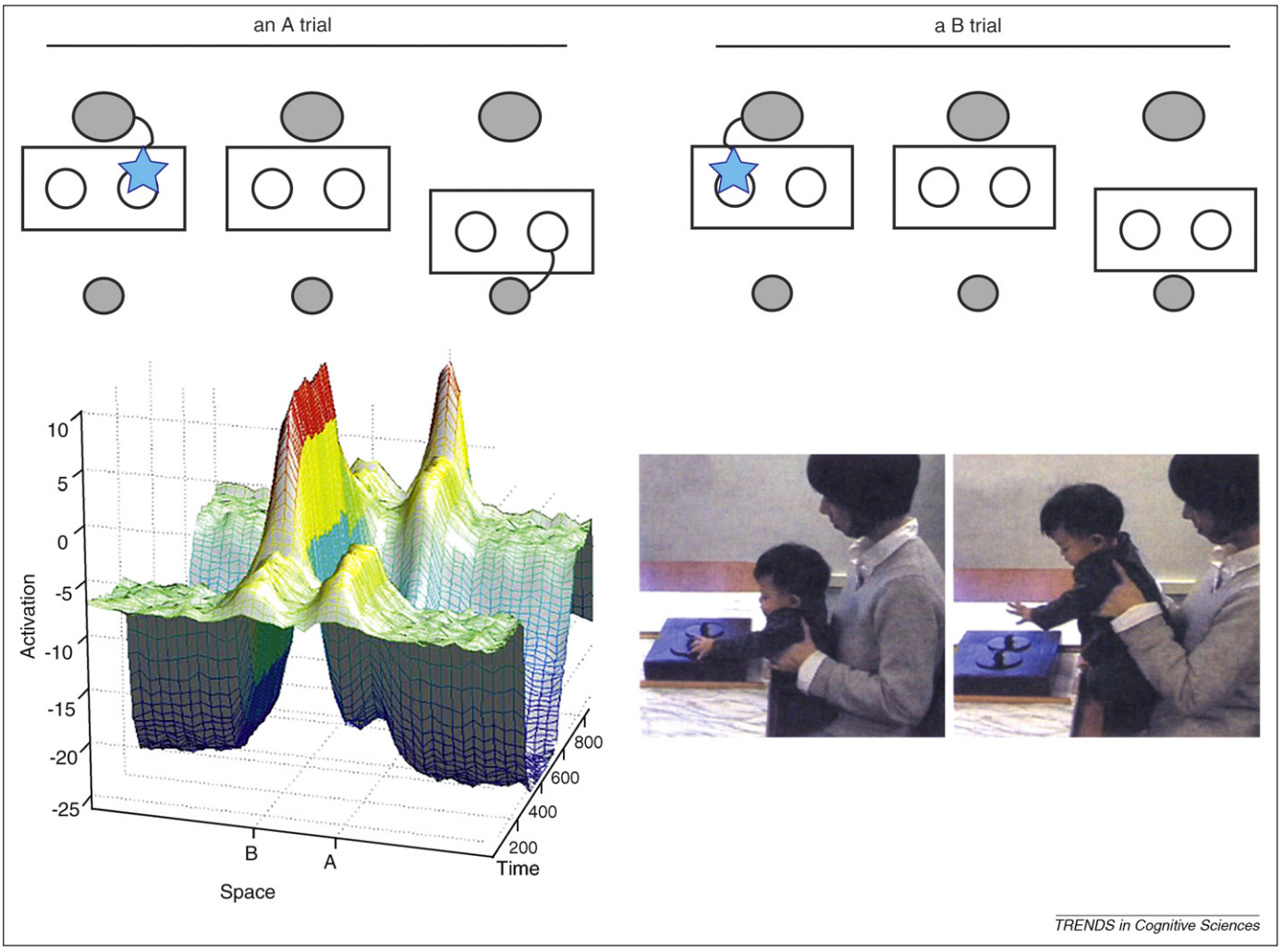

Figure 1. On the A trials, an experimenter hides an object repeatedly in one location, for example under a lid to the infant's right. The infant watches the object being hid, a delay of several seconds is imposed, and then the hiding box is pushed close to the infant and the infant is allowed to reach to the hiding location and retrieve the object. This is repeated several times: hiding under the rightmost lid, delay, infant retrieval of the object. On the crucial B trial, the experimenter hides the object in a new adjacent location, under a second lid to the infant's left. After the delay, the infant is allowed to reach. Bottom left: a DFT simulation of activation in the dynamic field on a B trial. The activation rises at the B location during the hiding event, but then, because of the cooperativity in the field and memory for previous reaches, activation begins to rise at A during the delay and the start of the reach inhibits the activation at B resulting in a simulated reach to A. Bottom right: a baby in a posture-shift A-not-B task.

independently of the infants' actions; younger infants, lacking such a hypothesis, reach to the place where their actions previously led them to find the object [11,12].

Experimental data favor an alternative, emergentist account of performance in the A-not-B task that has been developed within Dynamic Field Theory (DFT) [13,14]. This account explains the error through general processes of goal-directed reaching (and indeed is a variant of one model of adult reaching behavior). The model consists of a dynamic field, shown in Figure 1, which corresponds to the activation within a population of neuron-like units, each dynamically representing the direction of a reach. The field integrates multiple sources of relevant information: the immediate events (e.g. hiding the toy), the lids or covers on the table, and the direction of past reaches. The internal activations that produce a directional reach are themselves dynamic events, with rise times, decay rates, amplitudes and varying spatial resolution. Consequently, the model predicts – and experiments have confirmed – fine-grained stimulus, timing and task effects [13,14]. Because the explanation derives from general models of goal-directed action that are not specific to this task nor to this developmental period, the model makes predictions (tested and confirmed) about similar phenomena (and perseverations) at ages younger than, and considerably older than, the typical age range examined in the standard task [15,16]. Indeed, using this model as a guide, experimenters can make the error come and go predictably: by changing the delay, by heightening the attention-grabbing properties of the covers or the hiding event, and by increasing and decreasing the number of prior reaches to A [13,14,16,17].

The DFT-based model accounts for a wide range of findings showing that variables unrelated to beliefs about the existence of objects can affect the A-not-B error. The model has also been used to predict (correctly) that a reach back to A will occur in some situations when there is no toy hidden [17]. Furthermore, because the dynamic field is viewed as a motor planning field, and thus is tied to the body-centric nature of neural motor plans [17], the model also makes the novel prediction that perseverative errors should disappear if the motor plan needed for reaching to B is distinctly different from that for reaching to A [18]. One

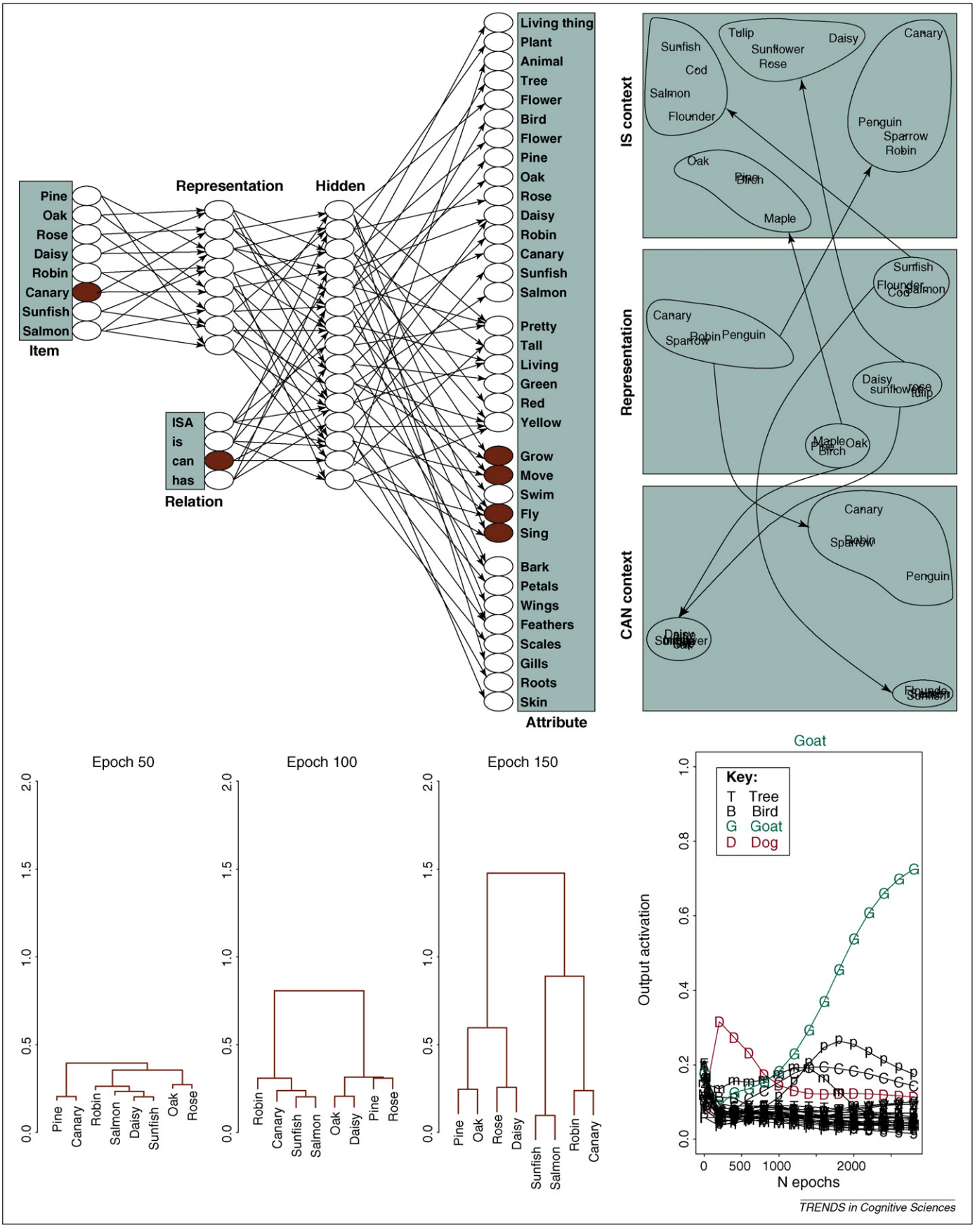

Figure 2. Top left: the connectionist network used by Rogers and McClelland [20], first used by Rumelhart and Todd [69], to explore the emergence of structure from experience. The network is trained by presenting item-context input pairs (e.g. 'canary' 'can') and then propagating activation forward (to the right) to activate units standing for possible completions of simple three-term propositions. Learning occurs by comparing the output to a pattern representing the valid completions (in this case, 'move'/'grow'/'fly'/'sing'), then adjusting connection weights throughout the network to reduce the discrepancy between the network's output and the valid completions. Learning occurs gradually, producing a differentiation progressive differentiation of items at the Representation layer and also influencing the patterns that emerge at the

experiment achieved this by shifting the posture of the infant ([17,19]; Figure 1).

Because the error can occur even when no object is hidden and can disappear with changes to the infant's posture, explanations based on beliefs about objects seem largely irrelevant to understanding A-not-B behavior. What is developing is a complex dynamic system, and it is this system that governs intelligent behavior, not the concepts, hypotheses or inferences that some ascribe to the infant's thinking.

Connectionist vs. structured probabilistic approaches to semantic cognition

We consider next a domain that both approaches have addressed, that of semantic cognition. Under the structured probabilistic approach [9], the acquisition of semantic knowledge is viewed as the inductive problem of deciding which of several alternative conceptual structures is most likely to have generated the observed properties of a set of items in a domain. This computation requires specification of considerable initial knowledge: (i) knowledge of the hypothesis space, the space of possible concepts and structures for relating concepts; (ii) prior distributions over the concepts and the structures. A similar approach has been taken to characterizing language acquisition [8].

Our fundamental disagreement with this approach concerns the fact that the alternative structured representations over which a probabilistic choice must be made generally do not, and perhaps cannot, adequately capture real-world domain structure [20]. For example, a hierarchical taxonomic model that has been fit to natural kinds [9] fails to take account of the presence of partial homologies across separate branches of the hierarchy, such that predatory birds, fish and mammals tend to share one set of properties whereas the prey of each kind tend to share others. Although the assignment of parallel structures might capture the strict homology, partial homologies would have to be force-fit. Similarly, a context-free grammar could provide a better fit to a corpus of sentences than some alternatives [8], but such grammars miss subtler probabilistic dependencies easily captured in connectionist models [21,22].

Connectionist models take a fundamentally different approach: the task of the model is not to choose from a set of prespecified alternative structures, but to learn a set of real-valued weights on connections among neuron-like processing units that support the generation of appropriate, context-sensitive, conditional expectations. Discrepancies between predicted and observed outcomes provide feedback for learning, in the form of gradual weight adjustment (Figure 2). Related items tend to evoke similar internal representations, thereby supporting generalization, although the system can use context to learn different similarity relations among the same sets of items when appropriate [20]. Similar approaches are used in connectionist models of semantic learning and language acquisition [21,22].

Although the continuous space of possible weight sets for a given connectionist network could be seen as analogous to the 'hypothesis space' of the structured probabilistic approach, there are several key differences. First, unlike the structured probabilistic approach [9], there is no restriction to a set of possible structure types, so that structures that do not exactly match any idealized type can be represented. Second, there is never a discrete decision to select one structure over another: the network's current set of weights can approximate one structure or a blend of structures. Third, learning simply involves the gradual refinement and elaboration of knowledge based on each new experience, and thus is far more constrained than the arbitrarily complex computation typically allowed by structured probabilistic approaches for computing the optimal structure from the entire corpus of relevant experiences.

A final point of comparison concerns inductive biases that play a role in both approaches. Whereas the hypothesis spaces of the structured probabilistic approach impose both general and domain-specific (content-based) biases, work within the connectionist approach has typically focused on the discovery of structure using only domain-general biases derived from properties of the learning procedure and network architecture [7,20]. Although content-based constraints can be built into connectionist models, connectionist work has focused on generic constraints that foster the discovery of structure, whatever that structure might be, across a range of domains and content types [7,20]. Yet, despite using only domain-general constraints, the connectionist model of semantic learning [20] explains evidence others [23,24] use to argue that children rely on innate domain-specific constraints. The model can acquire domain-specific patterns of responding: it can rely, for example, on shape over color for semantic judgments in one domain but on color over shape in another (see also [25]). Similar to children [26], the model can exploit different types of similarity among the same set of items in different contexts (e.g. taxonomically-defined similarity for biological properties, but a one-dimensional similarity space for judgments about size; Figure 2). The model also exhibits patterns of conceptual change that mirror phenomena reported in the literature, including: (i) a progressive differentiation in development (Figure 2), (ii) the advantage of basic level concepts in many situations but (iii) the elimination of the basic-level advantage in expertise, (iv) transient overgeneralization and illusory correlations in development and (v) the progressive disintegration of semantic knowledge in semantic dementia [27,28]. Models cast at a competence level have not addressed most of these phenomena.

Hidden layer, where representations are shaded by context. As learning progresses over successive sweeps through the set of item-context-output training patterns, the network first differentiates the plants from the animals and later differentiates the different types of animals and different types of plants. Upper right: the middle panel shows the similarity structure learned among a larger set of items in the Representation layer. The flanking panels show how this structure is reorganized in different contexts across units in the Hidden layer. Note that in the 'can' context, the plants are all represented as similar because they all do the same thing (they grow). Bottom right: naming response of the network when the input is 'goat' at different points in training. Note the transient tendency to activate 'dog' before the correct response 'goat' is acquired. In this instance, the network was trained in an environment where dogs were more frequent than other types of animal. Before 'dog' is differentiated from other animal types, the network treats all animals the same, naming them all with the most common animal name 'dog'. As differentiation occurs the correct name of 'goat' is finally learned. All panels reproduced, with permission, from Ref. [20].

Box 4. Outstanding questions

- What types of network architectures best promote the discovery of structure?
- To what extent are generic constraints sufficient to enable acquisition of domain-specific structure?
- When does the advantage of imposing a specific structural form on knowledge outweigh the disadvantages? Does expertise increase or decrease conformity to specific structural forms?
- When do humans truly engage in explicit hypothesis selection, and how can we distinguish such cases from situations in which they are gradually adapting implicit forms of knowledge such as connection weights in response to experience?

In short, the need to select among a prespecified set of alternative structure types in [9] forces semantic representation into an ill-fitting procrustean bed; the connectionist model of semantic cognition shows that this is unnecessary. Although further development of this model will certainly be required (Box 4; [29]), the model in its current form already shows that conceptual knowledge can emerge from a constrained learning process, without prior domain-specific knowledge and without requiring prespecification of possible knowledge structures or selection among them.

Conclusion

Far from being functionally equivalent or simply different levels of description, different theoretical frameworks lead to different conclusions about the nature of cognitive development, the kind of questions that a cognitive theory should address, and how explanations of different domains of behavior should be unified. The structured probabilistic approach takes the stand that it is crucial to specify the goal of cognitive processes at an abstract, computational or competence level of analysis before it makes sense to be concerned with the performance characteristics of particular algorithms or hardware implementations. Although this stance does not preclude explicit implementation, the properties of the machinery that implements the computations are not considered theoretically relevant. By contrast, the emergentist approach to understanding cognition, exemplified by dynamical systems and connectionist models, emphasizes the importance of specifying the actual mechanisms that underlie human cognitive performance, ultimately in terms of their neural implementation. The latter approach welcomes consideration of more abstract levels of description, and numerous research efforts have benefited considerably from integrating theories across levels [30,31], but not at the expense of mechanism (Box 5).

The commitment to mechanism is both principled and pragmatic. On the principled side, cognitive processing emerges out of evolutionary and developmental pressures and constraints that include the limited capabilities of biologically realizable hardware and the real-time demands of the environment. For example, biological vision could not have evolved solely as an in-principle response to the abstract problem of seeing because it was also constrained by what could evolve from previsual biological precursors and it had to operate in real time. Thus, the fundamental nature of cognitive processing is shaped by the performance characteristics of the underlying mechanism, and approaches that abstract away from such information run a serious risk of missing critical aspects of the problem under consideration.

On the pragmatic side, attention to both the strengths and limitations of specific implementation details has led to valuable theoretical advances that would have been unavailable if operating only at a competence level of analysis. A clear case in point concerns the observation that distributed connectionist networks suffer 'catastrophic interference' to old knowledge when forced to rapidly learn new inconsistent knowledge without the

Box 5. Emergentist approaches address function and mechanism: response to Griffiths *et al.* [2]

We view Griffiths and colleagues' arguments for their top-down, structured probabilistic models approach and against our emergentist one as misguided in at least three important respects.

The characterization of our view

The authors suggest that whereas their approach is 'top-down' ours is 'bottom-up.' Actually, we emphasize function, algorithm and implementation equally and seek accounts that span levels. We use dynamical systems and connectionist networks because they provide tools to address questions at all of these levels, including function. The 'function-first' approach will go astray if it makes incorrect assumptions about what the functions and goals actually are. In fact, we question many of their assumptions about function: for example, that the goal of language acquisition is to induce grammatical rules, or that the goal of semantics is to induce a structure representing relations among concepts. If these are not the right problems, the question of how to solve them optimally is moot. Mechanistic commitments place important constraints on the kinds of computations that are easy or natural, and thus provide information about what functions are actually computed. Thus, attention to mechanism can provide clues to function and attention to function can provide clues to mechanism.

The characterization of human abilities

The authors assume that human behavior is rational, and that cognition is compositional and recursive. In so doing they seem to overestimate and mischaracterize human cognitive abilities. For instance, they suggest that people can radically reconfigure their beliefs on the basis of a single statement—as though hearing a phrase like ''dolphins are not fish but mammals'' will dramatically reorganize the listener's knowledge about animals. Although people can memorize arbitrary facts, deep conceptual reorganization occurs gradually over years, and coexists with knowledge of inconsistent facts. Human behavior is also notoriously susceptible to biases and heuristics that can lead to violations of rationality. To the extent that such behaviors can be explained post hoc by 'rational' models, the models are underconstrained: any pattern of human behavior will be consistent with some rational analysis of the problem. To be useful, a theoretical account must explain not only why people excel at some cognitive abilities but also why they fail at others.

The characterization of the capacities of emergentist models

Several of Griffiths *et al.*'s statements about the limitations of emergentist models are incorrect. Contra their statements, such models can: (i) exploit information provided by natural language or social context [64], (ii) account for rapid learning and generalization of new words [34,65], (iii) explain why people sometimes generalize in an all-or-none fashion and sometimes in a graded fashion [66], (iv) explain nonlinearities in children's lexical development [67,68], and (v) explain why people generalize differently in different contexts [20]. Although emergentist models are constrained in what they can do easily, we view this as an advantage. The constraints arise from a commitment to mechanisms similar to those that implement real minds, thus they provide useful clues as to how real minds solve important cognitive problems.

chance to rehearse the old knowledge [32,33]. Such rapid learning is possible using very sparse representations, but this compromises the ability to learn the underlying statistical structure of experiences, thereby undermining generalization. The competing demands of rapid learning of new knowledge versus the gradual discovery of underlying structure are consequences of the connectionist implementation of learning and memory. This competition led McClelland, McNaughton and O'Reilly [34] to propose that these functions are subserved by distinct but complementary memory systems – hippocampus and neocortex, respectively – with the former helping to consolidate knowledge in the latter over time. There are other possible implementations of mechanisms of learning and memory in which there is no conflict between these demands. Thus there is no basis for understanding the contrasting properties and coordinated operation of hippocampus and neocortex without committing to properties of the mechanism.

In summary, we advocate an integrated approach to cognition in which functional considerations are grounded in, and informed by, the performance characteristics of the underlying neural implementation.

References

1 Johnson, S.B. (2001) *Emergence: The connected lives of ants, brains, cities, and software*, Scribner's
2 Griffiths, T.L. *et al.* (2010) Probabilistic models of cognition: Exploring the laws of thought. *Trends Cogn. Sci.* 14, 357–364
3 Marr, D. (1982) *Vision*, W. H. Freeman
4 Chomsky, N. (1965) *Aspects of the theory of syntax*, MIT Press
5 Sternberg, D. and McClelland, J.L. (2009) When should we expect indirect effects in human contingency learning? In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Taatgen, N.A. and van Rijn, H., eds), pp. 206-211, Cognitive Science Society
6 Movellan, J.R. and McClelland, J.L. (1993) Learning continuous probability distributions with symmetric diffusion networks. *Cogn. Sci.* 17, 463–496
7 Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the dimensionality of data with neural networks. *Science* 313, 504–507
8 Perfors, A. *et al*. (2006) Poverty of the stimulus? A rational approach. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 663–668, Lawrence Erlbaum Associates
9 Kemp, C. and Tenenbaum, J.B. (2009) Structured statistical models of inductive reasoning. *Psychol. Rev.* 116, 20–58
10 Sobel, D. *et al.* (2004) Children's causal inferences from indirect evidence: backwards blocking and Bayesian reasoning in preschoolers. *Cogn. Sci.* 28, 303–333
11 Piaget, J. (1954) *The construction of reality in the child*, Basic Books
12 Baillargeon, R. (1994) How do infants learn about the physical world? *Curr. Dir. Psychol. Sci.* 3, 133–140
13 Thelen, E. *et al.* (2001) The dynamics of embodiment: A field theory of infant perseverative reaching. *Behav. Brain Sci.* 24, 1–86
14 Clearfield, M.W. *et al.* (2009) Cue salience and infant perseverative reaching: Tests of the dynamic field theory. *Dev. Sci.* 12, 26–40
15 Clearfield, M.W. *et al.* (2006) Young infants reach correctly in A-not-B tasks: On the development of stability and perseveration. *Infant Behav. Dev.* 29, 435–444
16 Spencer, J.P. *et al.* (2001) Tests of a dynamic systems account of the A-not-B error: The influence of prior experience on the spatial memory abilities of two-year-olds. *Child Dev.* 72, 1327–1346
17 Smith, L.B. *et al.* (1999) Knowing in the context of acting: The task dynamics of the A-not-B error. *Psychol. Rev.* 106, 235–260
18 Diedrich, F.J. *et al.* (2000) Motor memory is a factor in infant perseverative errors. *Dev. Sci.* 3, 479–494
19 Lew, A. *et al.* (2007) Postural change effects on infants' AB task performance: Visual, postural, or spatial? *J. Exp. Child Psychol.* 97, 1–3
20 Rogers, T.T. and McClelland, J.L. (2004) *Semantic cognition: A parallel distributed processing approach*, MIT Press
21 Elman, J.L. (1990) Finding structure in time. *Cogn. Sci.* 14, 179–211
22 Rohde, D.L.T. and Plaut, D.C. (1999) Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition* 72, 67–109
23 Keil, F.C. (1981) Constraints on knowledge and cognitive development. *Psychol. Rev.* 88, 197–227
24 Gelman, R. (1990) First principles organize attention to and learning about relevant data: Number and the animate/inanimate distinction as examples. *Cogn. Sci.* 14, 79–106
25 Colunga, E. and Smith, L.B. (2005) From the lexicon to expectations about kinds: a role for associative learning. *Psychol. Rev.* 112, 347–382
26 Gelman, S.A. and Markman, E.M. (1986) Categories and induction in young children. *Cognition* 23, 183–209
27 Rogers, T.T. *et al.* (2004) The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychol. Rev.* 111, 205–235
28 McClelland, J.L. *et al*. (2009) Semantic cognition: Its nature, its development, and its neural basis. In *The cognitive neurosciences IV* (Gazzaniga, M., ed.), pp. 1047–1066, MIT Press
29 Rogers, T.T. and McClelland, J.L. (2008) A simple model from a powerful framework that spans levels of analysis. *Behav. Brain Sci.* 31, 729–749
30 Botvinick, M.M. and An, J. (2008) Goal-directed decision making in prefrontal cortex: A computational framework. In *Advances in Neural Information Processing Systems* (Koller, D. *et al.*, eds), pp. 169–176, Curran Associates, Inc.
31 McClelland, J.L. and Chappell, M. (1998) Familiarity breeds differentiation: A subjective-likelihood approach to the effeccts of experience in recognition memory. *Psychol. Rev.* 105, 724–760
32 McCloskey, M. and Cohen, N.J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In *The psychology of learning and motivation* (Bower, G.H., ed.), pp. 109–165, Academic Press
33 Ratcliff, R. (1990) Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychol. Rev.* 97, 285–308
34 McClelland, J.L. *et al.* (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. *Psychol. Rev.* 102, 419–457
35 Feldman, N.H. *et al.* (2009) The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychol. Rev.* 116, 752–782
36 Hay, J. (2001) Lexical frequency in morphology: Is everything relative? *Linguistics* 39, 1041–1070
37 Bybee, J. and McClelland, J.L. (2005) Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *Ling. Rev.* 22, 381–410
38 Bybee, J.L. (1985) *Morphology: A study of the relation between meaning and form*, John Benjamins
39 Gonnerman, L.A. *et al.* (2007) A distributed connectionist approach to morphology: Evidence from graded semantic and phonological effects in lexical priming. *J. Exp. Psychol. Gen.* 136, 323–345
40 Culicover, P.W. (1999) *Syntactic nuts: Hard cases in syntax. Volume 1, Foundations of Syntax*, Oxford University Press
41 Prince, A. and Smolensky, P. (2004) *Optimality theory: Constraint interaction in generative grammar*, Blackwell
42 Smolensky, P. and Legendre, G. (2006) The harmonic mind: *From neural computation to Optimality-theoretic grammar, Vol. 1: Cognitive architecture, Vol. 2: Linguistic and philosophical Implications*, MIT Press
43 Rumelhart, D.E. *et al*. (1986) Schemata and sequential thought processes in PDP models. In *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2) (McClelland, J.L. *et al*., eds), pp. 7-57, MIT Press
44 Plaut, D.C. *et al.* (1996) Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115
45 Plaut, D.C. and Gonnerman, L.M. (2000) Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Lang. Cogn. Proc.* 15, 445–485

46 Pinker, S. and Ullman, M.T. (2002) The past and future of the past tense. *Trends Cogn. Sci.* 6, 456–463
47 Rumelhart, D.E. and McClelland, J.L. (1986) On learning past tenses of English verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2) (McClelland, J.L. et al., eds), pp. 216–271, MIT press
48 Joanisse, M.F. and Seidenberg, M.S. (1999) Impairments in verb morphology following brain injury: a connectionist model. *Proc. Natl. Acad. Sci.* 96, 7592–7597
49 Seidenberg, M.S. and McClelland, J.L. (1989) A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568
50 Elman, J.L. (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* 7, 195–224
51 Servan-Schreiber, D. *et al.* (1991) Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Mach. Learn.* 7, 161–193
52 St. John, M.F. and McClelland, J.L. (1990) Learning and applying contextual constraints in sentence comprehension. *Artif. Intell.* 46, 217–257
53 St. John, M.F. (1992) The story gestalt: A model of knowledge-intensive processes in text comprehension. *Cogn. Sci.* 16, 271–306
54 Jansen, B.R.J. and van der Maas, H.L.J. (2001) Evidence for the phase transition from Rule I to Rule II on the balance scale task. *Dev. Rev.* 21, 450–494
55 Ferretti, R.P. and Butterfield, E.C. (1986) Are childrens' rule-assessment classifications invariant across instances of problem types? *Child Dev.* 57, 1419–1428
56 McClelland, J.L. (1989) Parallel distributed processing: Implications for cognition and development. In *Parallel distributed processing: Implications for psychology and neurobiology* (Morris, R., ed.), pp. 8–45, Oxford University Press
57 Schapiro, A.C. and McClelland, J.L. (2009) A connectionist model of a continuous developmental transition in the balance scale task. *Cognition* 110, 395–411
58 Denny-Brown, D. (1966) *The cerebral control of movement*, Charles C. Thomas
59 Thelen, E. and Smith, L.B. (1994) *A dynamic systems approach to the development of cognition and action*, MIT Press
60 Farah, M.J. and McClelland, J.L. (1991) A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *J. Exp. Psychol. Gen.* 120, 339–357
61 Schwartz, M.F. *et al.* (1991) The quantitative description of action disorganization after brain damage: A case study. *Cogn. Neuropsychol.* 8, 381–414
62 Botvinick, M. and Plaut, D.C. (2004) Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychol. Rev.* 111, 395–429
63 Botvinick, M. and Plaut, D.C. (2006) Short-term memory for serial order: A recurrent neural network model. *Psychol. Rev.* 113, 201–233
64 Hirsh-Pasek, K. *et al*. (2000) An emergentist coalition model for word learning: mapping words to objects is a product of the interaction of multiple cues. In *Becoming a Word Learner: a Debate on Lexical Acquisition* (Hirsh-Pasek, K. and Golinkoff, R.M., eds), pp. 136-164, Oxford Press
65 Mayor, J. and Plunkett, K. (2010) A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychol. Rev.* 117, 1–31
66 McClelland, J.L. and Patterson, K. (2002) Rules or connections in past-tense inflections: What does the evidence rule out? *Trends Cogn. Sci.* 6, 465–472
67 Marchman, V. and Bates, E. (1994) Continuity in lexical and morphological development: A test of the critical mass hypothesis. *J. Child Lang.* 21, 339–366
68 MacWhinney, B. (1998) Models of the emergence of language. *Ann. Rev. Psychol.* 49, 199–227
69 Rumelhart, D.E. and Todd, P.M. (1993) Learning and connectionist representations. In *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (Meyer, D.E. and Kornblum, S., eds), pp. 3–30, MIT Press