5-1-2015

Connectionist perspectives on language learning, representation and processing.

Marc F Joanisse
Psychology/Brain and Mind Institute, The University of Western Ontario, London, ON, Canada

James L McClelland
Department of Psychology, Stanford University, Stanford, CA, USA

Advanced Review

Connectionist perspectives on language learning, representation and processing

Marc F. Joanisse[1]* and James L. McClelland[2]

The field of formal linguistics was founded on the premise that language is mentally represented as a deterministic symbolic grammar. While this approach has captured many important characteristics of the world's languages, it has also led to a tendency to focus theoretical questions on the correct formalization of grammatical rules while also de-emphasizing the role of learning and statistics in language development and processing. In this review we present a different approach to language research that has emerged from the parallel distributed processing or 'connectionist' enterprise. In the connectionist framework, mental operations are studied by simulating learning and processing within networks of artificial neurons. With that in mind, we discuss recent progress in connectionist models of auditory word recognition, reading, morphology, and syntactic processing. We argue that connectionist models can capture many important characteristics of how language is learned, represented, and processed, as well as providing new insights about the source of these behavioral patterns. Just as importantly, the networks naturally capture irregular (non-rule-like) patterns that are common within languages, something that has been difficult to reconcile with rule-based accounts of language without positing separate mechanisms for rules and exceptions.

How to cite this article:
WIREs Cogn Sci 2015. doi: 10.1002/wcs.1340

INTRODUCTION

Formal approaches to linguistics following Chomsky's Generative Grammar framework envision language representation as an assembly of symbols (e.g., words and phrases) and a grammar of rules that operates on them.[1–3] Questions in formal linguistics have thus tended to focus on the correct formalization of the rules, what their scope is, as well as how they are learned. In this review, we discuss a completely different approach to thinking about language representations, from the parallel distributed processing (PDP) or 'connectionist' point of view. This perspective eschews the concepts of symbols and rules in favor of a model of the mind that closely reflects the functioning of the brain. As we will discuss, this approach allows us to account for a wide range of linguistic data using a much more restricted set of assumptions. We begin with a brief overview of the connectionist enterprise and its basic assumptions about how mental processes can be studied using networks of artificial neurons.

*Correspondence to: marcj@uwo.ca

[1]Psychology/Brain and Mind Institute, The University of Western Ontario, London, ON, Canada

[2]Department of Psychology, Stanford University, Stanford, CA, USA

Conflict of interest: The authors have declared no conflicts of interest for this article.

OVERVIEW OF CONNECTIONISM

The connectionist approach to language builds on some key guiding assumptions about the nature of mental representations[4,5]:

1. *Knowledge is represented as patterns of numerical activity across large sets of simple processing units*: Mental states reflect the activation of neurons in the brain. These patterns are *distributed*

such that knowledge of individual concepts or categories occurs through the activation of many individual processing units. Likewise, no single neuron uniquely encodes a concept or category; rather individual neurons can be re-used to encode many different concepts.

2. *Processing occurs via transformations of patterns of activity across large sets of connections*: Neurons in the brain are massively interconnected. This allows information to be retrieved and processed by transforming activity among large assemblies of artificial neurons.
3. *Learning occurs as the confluence of innate but domain–general architectural and learning mechanisms, plus experience*: Networks learn via changes in the strength of connections among interconnected units, in response to external inputs (the environment). This process is governed by general laws of learning that are not specific to any single type of process. Just as importantly, these neurons are not organized haphazardly. Rather they have distinct biologically specified architectural characteristics that also influence how learning proceeds.[a]

A central component of the connectionist enterprise is to develop computational simulations of key phenomena. This allows us to make explicit assumptions about the nature of the processes and representations of interest. Implementing these into a model then provides an explicit test of these assumptions, as well as a way to test hypotheses about them. In addition, the results of models provide new hypotheses that can be tested empirically in humans.

Connectionist models encompass a number of simplifying assumptions that abstract away from actual brains in some important ways; specifically they tend to contain many fewer processing units than what one finds in the brain. In addition, these models are made up of artificial neurons that represent rates of neural firing as static activation levels, which change in response to inputs from the environment and from other units. Finally, the learning mechanisms tend to be computationally simpler than those that we know govern actual learning in neurons. The purpose of these simplifying assumptions is to create models that capture the assumptions laid out above, while keeping the model sufficiently simple so as to be implemented within a computer program.

Connectionism Applied to Language

The connectionist enterprise was conceived as a way to address a wide range of cognitive phenomena. Just as importantly, it is seen as a *unifying* theory, because it assumes all types of mental knowledge can be understood within it. Thus, it does not assume a strong distinction between language and other types of knowledge. In this sense, connectionism is in conflict with some of the guiding assumptions of the generative linguistics framework, which has historically built on the idea that language is learned and represented using mechanisms that are distinct from those governing other types of knowledge.

Formal linguistics itself grew out of a concern about the learnability of language and the need to establish innate language-specific mechanisms of learning.[7] Consequently, the re-emphasis that connectionism places on these concepts might appear like a regression of sorts. That said, we argue here that connectionist mechanisms are able to learn and encode complex knowledge in ways that are not trivial. As we will show, connectionist models can capture the rule-like patterns that are observed in language. Likewise, the patterns of learning observed in these models also can closely resemble the way that children learn language. And importantly, the models are able to also capture irregular (non-rule-like) patterns that also pervade languages without requiring a separate mechanism to do so.

AUDITORY WORD RECOGNITION

An important contribution of connectionist theories of language processing has been the idea of dynamicism and interactivity in language processing. That is, these models lend themselves well to processes that involve recognizing inputs through the interaction of bottom–up sensory information and top–down contextual/experiential information. Here we consider one such phenomenon, that of spoken word recognition. This task can be seen as a 'hard problem' in language processing; the listener must rapidly segment individual words from a connected spoken utterance, and then identify them from among many different competing forms. This is especially difficult given variability in the acoustic cues of individual phonemes (e.g., effects of coarticulation in which a phoneme is realized differently depending on what other phonemes precede and follow it).

Earlier models of auditory word recognition abstracted away from these issues by characterizing the task as one of lexical access. On this view, a sensory input is first broken down into its constituent phonemes, and these are then used to search for a discrete entry within a mental lexicon.[8] This is seen as a serial process in which the input is compared against all known lexical forms until a matching form

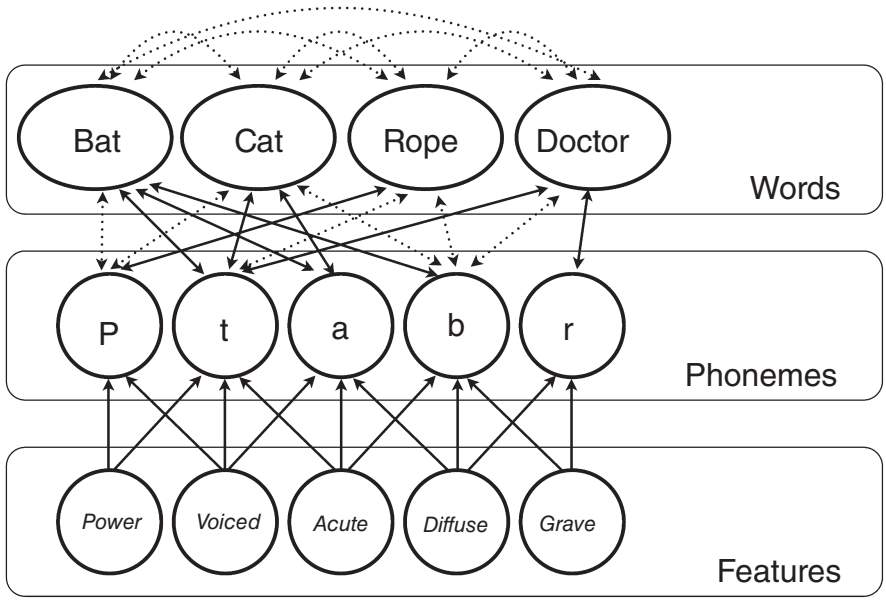

FIGURE 1 | The TRACE model of auditory word recognition.

is identified (e.g., Ref 9). Importantly, this type of serial search is proposed to be independent of the perceptual mechanisms used to map acoustic inputs onto phonetic and phonemic features.

McClelland and Elman[10] set out an alternative view, in the form of the connectionist TRACE model.[b] The model consists of three layers of neurons, used to represent auditory, phonemic, and word-specific information (Figure 1). It simulates word recognition by taking input as a time-varying acoustic–phonetic representation of a word, which in turn activates the word's corresponding phonemes and ultimately a single word-level unit that uniquely identifies it. So, for example, the word BAT is recognized by presenting as an input a sequence of acoustic–phonetic patterns that correspond to this word (i.e., a numerical activity pattern corresponding to the presence or absence of different phonetic features in each of its phonemes); activation then propagates to the phonemic layer in order to activate individual units that encode the phonemes /b/, /æ/, and /t/, and finally a single word-level unit that uniquely represents the concept 'bat'.

A key characteristic of the model is its *dynamical* nature. The network receives an auditory input that changes over time, rather than all the information at once. Time is divided into discrete processing 'cycles' in which activation is propagated from one layer of neurons to another. Thus, words are recognized incrementally by slowly ramping up the activation of the correct units at the phoneme and word levels. Critically this is different from a serial search model in which a model must search through individual lexical entries one at a time until the correct one is found. Here all forms compete for selection in parallel. The model is also *interactive*; it contains connections that project both bottom–up and top–down, so that activation at the word-level can influence activation at the phoneme level. As we show below, this has important consequences for how the model processes information, especially in the case of ambiguous or missing inputs.

The model provides explanations to a range of phenomena in speech perception. In the interest of space, we focus here on the broad category of 'lexical' effects, which are concerned with word-level phenomena. TRACE emphasizes the role of top–down influences in speech processing acting on the phoneme layer as well as the feature layer. Feedback connections from the word layer to the phoneme layer allow the model to supplement bottom–up sensory information with top–down word-specific information, which has a number of desirable consequences. Take for instance the phoneme restoration effect[11]: when one of a word's phonemes is replaced with a burst of white noise (e.g., the /s/ in 'legislature'), listeners nevertheless report hearing the missing sound. So, for example, they have difficulty reporting whether a sound has been deleted and replaced by the noise or whether the noise has simply been added to the word, confirming that they are experiencing an auditory illusion in which the missing phoneme has been restored.

The phoneme restoration effect appears to occur because listeners use top–down information to

supplement the imperfect bottom–up sensory information. This can be simulated within TRACE by presenting the model with an incomplete acoustic input. For instance, /b#ek/ (the word 'break' but with the features of /r/ replaced with random noise denoted by the # symbol) still activates three of the four phoneme units corresponding to the target word. This in turn leads to partial activation of the correct word level unit. Words in this model have feedback connections that activate their constituent phonemes. In the case of the word 'break', this means that the /r/ phoneme unit can become activated through top–down activation from the word-level 'break' unit even when the bottom–up (perceptual) information is incomplete or incorrect. As a result, the model is able to 'repair' the input by activating a phoneme that was in fact missing from the input.

Top–down effects also allow the model to divide an unsegmented input stream like /barti/ into its two constituent words *bar* and *tea*. However, this tendency is weaker when the longer word is itself a familiar form. That is, the input /parti/ also tends to activate the words *par* and *tea*, albeit to a much lesser extent than the longer word *party*. Notably, these sorts of patterns fall naturally out of the dynamics of the model, due to the assumption in the model that word units compete with each other to the extent that they encompass overlapping portions of the spoken input.

Lexical effects can also interact with sublexical information in interesting ways. One well-studied finding is that listeners' categorization of an ambiguous phoneme can be biased toward producing familiar words. For example, while listeners show the usual categorization profile for a VOT continuum between the phonemes /d/ and /t/ in isolation, categorization profiles tend to shift if these are presented in the context of a familiar carrier word. For instance, presenting an alveolar stop with an ambiguous VOT in the context of '_ask' yields a subtle bias toward categorizing it as /t/ rather than /d/.[12] This occurs even when listeners are asked to identify the initial consonant, and ignore the word it is embedded in. This effect again falls naturally out of how TRACE identifies phonemes: partial inputs will activate the word-level representation of 'TASK', and this projects back to the /t/ unit within the phoneme layer, yielding a subtle but reliable shift in the model's categorization curve along the /t/-/d/ continuum (Figure 2).

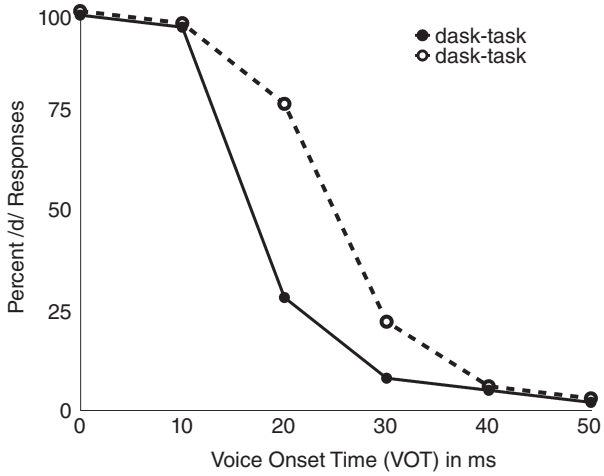

FIGURE 2 | Lexicality effect in phoneme categorization profile of the TRACE model. The categorization of a midpoint (ambiguous) stop consonant shifts as a function of the word in which it is embedded. As in humans, the model shows a preference toward real words over a nonword, but only when the phoneme's voicing parameter is near the category boundary.

Recent Developments

Although the TRACE model is now over 20 years old, interest in the model appears to be growing; the rate of citations of the original work have in fact increased since 2001. One reason for this is the growing corpus of behavioral studies examining the dynamics of spoken word recognition using eyetracking. Tanenhaus et al.[13,14] established a 'visual world' paradigm in which they present an array of visual objects as subjects hear words or sentences corresponding to it. They have found that listeners tend to show eye movements to the corresponding object starting about 200 ms after hearing its name (for an overview, see Ref 15). Strikingly, listeners also produce eye movements to objects corresponding to target words' phonological competitors. So, for instance, hearing the auditory word *candle* yields fixations to a picture of a candle, but also to pictures of 'candy' and 'sandal' (Figure 3(a)).[13]

What is notable is that both the timing and proportion of looks to target pictures and their phonological competitors closely matches what the TRACE model yields when presented with a similar task. That is, when the model is presented with *candle*, it also tends to activate phonological competitor words like *candy* and *sandal* (Figure 3(b)). Tanenhaus et al.[16] propose that this is no coincidence, and that there is a close link between eye movements to a given object and the activation of that word. As illustrated in Figure 3(b), this is captured within TRACE via differing degrees of activation of the competing words over time. Moreover, the model's dynamics closely match eye movement rates in humans. The target word is not selected immediately but instead shows activation ramping up over time. Concurrently, we see activation of competitor words rise and fall as a function of the degree to which they match the provided input. This

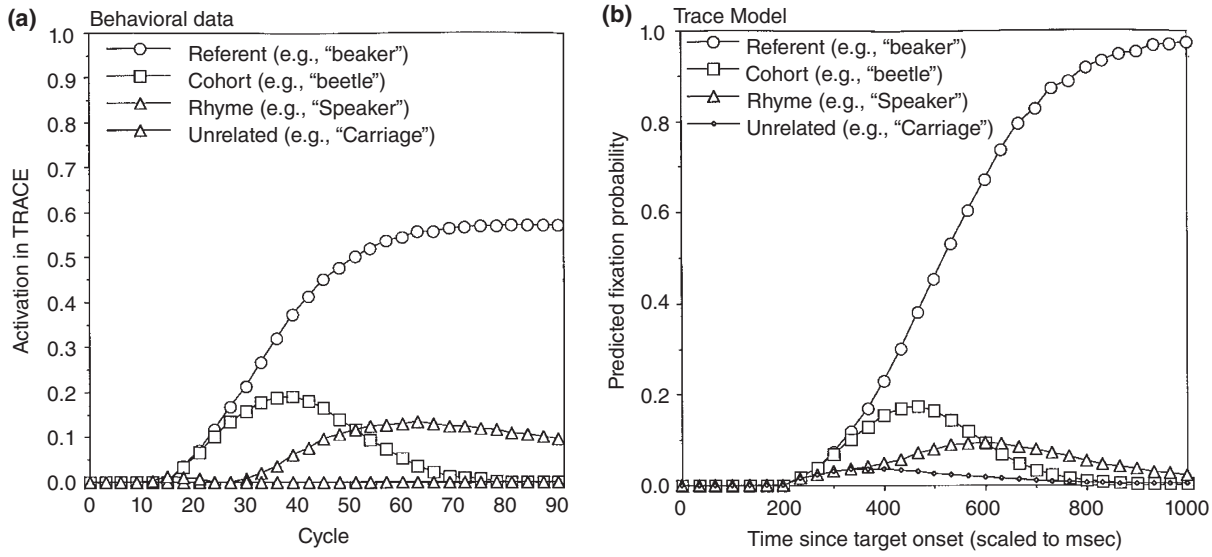

FIGURE 3 | Eye-tracking data showing competition effects from onset and rhyme competitors in a visual world paradigm. Both (a) adult listeners and (b) the TRACE model show comparable competition effects, marked by a larger proportion of eye movements to either type of phonologically related competitor relative to a phonologically unrelated foil. (Reprinted with permission from Ref 13. Copyright 2015 Elsevier).

includes an earlier effect of cohort competitors (words matching the initial phonemes; *candle–candy*) and a somewhat later going interference effect from rhyme competitors (*candle–sandal*).

The visual world paradigm has also been used to examine other types of lexical effects related to word frequency and phonological neighborhood density.[14,17] Here again, the data appear to closely match the predictions of TRACE, underlining the usefulness of the model in understanding the dynamics of auditory word recognition.

Controversy has surrounded the assumption in TRACE that lexical information really can feed back down to the phoneme level. Some have questioned the need for this, arguing that lexical influences can be taken into account in a postperceptual decision stage. One response to this has been to note that feedback to the phoneme level can have 'knock-on' effects, facilitating (1) the processing of neighboring phonemes or (2) the processing of subsequent tokens of the identified phoneme itself (see Ref 18, for a review). One specific example should serve to illustrate the general form of the argument. Suppose a listener encounters an individual with an unfamiliar dialect, in which some particular speech sound—say the /ʃ/ phoneme—is pronounced in a way that is unfamiliar to the listener. Lexical context may help the listener to identify this sound when it occurs in a context, such as /ful⌑#/ (i.e., the word *foolish* ending with novel instance of /ʃ/). If activation then flows top–down to cause the activation for the phoneme unit for /ʃ/, this could trigger the adjustment of the incoming connections to the /ʃ/ unit, so that next time the same input will activate /ʃ/ more strongly, adapting the listener's network to the speaker's unfamiliar dialect.

TRACE has also influenced how researchers understand word recognition processes in bilinguals. One of the key questions in this field concerns the extent to which bilingual speakers maintain separate phonological and/or lexical representations of their two languages. An early model of this is Grosjean's BIMOLA model,[19] which proposed separate parallel phonological and lexical layers for two languages, receiving inputs from a common shared feature layer. This model proposes that listeners will use the acoustic inputs to activate both sets of layers in parallel, and select the correct word based on which generates the strongest output among the two languages.

Dijkstra and Van Heuven[20] have proposed a competing model of bilingual word processing that proposes a much weaker division between the two competing languages. Their model deals specifically with reading rather than spoken word recognition, but nevertheless builds on the same principles of interactive activation as TRACE and BIMOLA. It proposes that words of both languages are maintained within a single mechanism, and are held separate in processing thanks to top–down connections from language-level nodes. A potential benefit of this model is the ability to account for the finding that bilingual individuals tend to show activation of competitor words across languages. For instance, French/English bilinguals show crossmodal priming effects such as faster recognition for the word BREAD when it is preceded

by the word PAIN (which is French for *bread*).[21] Such findings suggest that words in the two languages are not being held completely separately during processing. As the literature on bilingual word processing develops it is interesting to see how competing connectionist models can help adjudicate among different theories of how two languages are represented in one mind. Work on TRACE continues, spurred by the release of a new implementation of the model that can be run on modern computers (jTRACE).[22] Also noteworthy is a proposal from Hannagan et al.[23] of how the TRACE model might better capture the temporal nature of speech. The original network simulated the temporal nature of the speech signal within a static input scheme that presented different time points concurrently. By more accurately capturing how spoken words unfold over time, this refinement might provide even more fine-tuned insights into speech perception phenomena.

READING AND PAST TENSE

Perhaps the best-known connectionist models of language have focused on related phenomena of visual word recognition and past tense morphology. While these models have arisen to deal with somewhat different phenomena in psycholinguistics, the facts are similar across the two. Specifically, they are concerned with the mechanisms by which we acquire and process the regularities in language in parallel with the exceptional cases that also occur. Consider the case of past tense in English: a large majority of verbs (about 87%) are marked as past tense by adding a variant of the -ed suffix (*walk–walked* and *need–needed*). The ending is also highly productive, such that novel forms nearly always take the -ed form. Thus, listeners typically judge that a nonword form like *wug* or a neologism like *blog* will take a regular ending as in *wugged* and *blogged*. On this basis, it is assumed that regular past tense verbs are not stored outright, but rather are produced using a generative rule that transforms a verb stem into a past tense form by concatenating the -ed suffix. However, there are also a number of irregular verbs in English that defy such a rule (e.g., *take–took*, *sleep–slept*, and *go–went;* cf. **taked, *sleeped*, and **goed*). A popular theory has posited separate cognitive mechanisms for applying a rule to regular forms, and for memorizing word-specific knowledge of irregular forms.[24]

Reading in English presents a similar challenge. The mapping from print to sound is generally regular. For instance, words that begin with the letter B usually also begin with the /b/ phoneme; similarly, words that end in AVE tend to rhyme with each other (GAVE, SAVE, RAVE, CAVE, and PAVE). These regularities can then generalize to nonwords (MAVE) and neologisms (BLOG), which suggests readers encode regularities within a productive mechanism. That said, English spelling is rife with exceptions (e.g., HAVE does not rhyme with CAVE; the W in SWORD is silent; THROUGH, ROUGH, and DROUGHT are all pronounced differently).

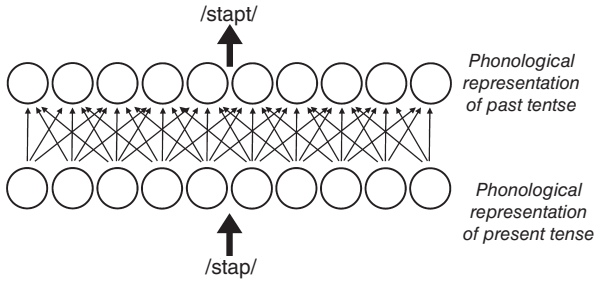

FIGURE 4 | The Rumelhart and McClelland's[25] model of past tense.

How do we handle both the productive and exceptional aspects of language? The answer from the connectionist standpoint is a single distributed mechanism that encodes both within a single network of connections. With respect to past tense, Rumelhart and McClelland[25] (RM86) proposed a model designed to learn the mapping between a verb's present and past tense forms. The model receives as input the phonological form of a present tense verb and has to produce, as an output, the verb's past tense (Figure 4). The model is trained on a representative corpus of English monosyllabic verbs that includes both regular and irregular forms. It uses a learning algorithm that adjusts connection weights based on experiences with correct past tense forms, such that the model's performance gradually improves with experience.

The resulting network is able to learn both regular and irregular forms within the same architecture, without appealing to the concepts of either 'rules' or 'memorized lexical entries'. The model also shows good generalization to novel forms, suggesting it is can take advantage of similarities in English past tense (e.g., given *blug,* it can produce *blugged*).

Seidenberg and McClelland[26] (herein, SM89) have proposed a similar approach to understanding visual word recognition. Word knowledge is modeled as the confluence of orthographic, phonological and semantic codes, each encoded within separate layers of a network (Figure 5). Learning to read involves learning the mapping among these three types of knowledge. Their original instantiation focused specifically on mapping orthography to phonology; their model was presented with the orthographic form of an English word as input, and learned to output its phonological form. The model was trained on a

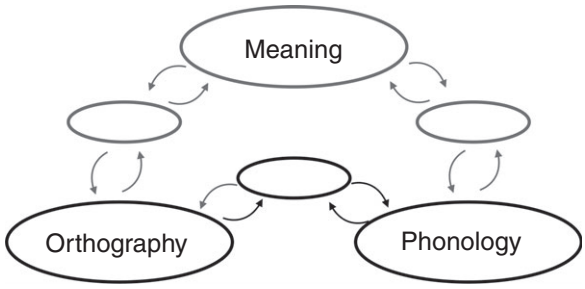

FIGURE 5 | The Seidenberg and McClelland's[26] model of reading. Portions in black depict the model as it was originally implemented.

corpus of several hundred English words, providing it with experience with both regular and irregular words. One important innovation was the use of a frequency-weighted training vocabulary. In this approach, the model was presented with different words at different rates, in a way that reflects the statistical properties of English. This was in turn reflected in the model's connection strengths for different types of patterns.

The fully trained network showed a number of desirable patterns of behavior: it tended to have greater difficulty with irregulars than regulars, marked by somewhat higher unit-wise differences between the desired and actual output values (quantified as 'sum-squared error' or SSE, which can be conceptualized as the difficulty it has in computing the correct output). Word frequency also influenced learning such that lower frequency forms tended to be more difficult to produce (reflected in higher SSEs). Finally, the model showed a frequency by regularity interaction in which the highest SSEs were observed for low frequency irregulars compared to high frequency irregulars and both high and low frequency regulars. This pattern closely resembles skilled readers' reaction times for similar words, and suggests the model is accurately capturing the cognitive mechanisms used in visual word recognition.

As discussed above, both the reading and past tense models appear to learn regular and irregular forms in parallel, and show output patterns consistent with what we observe in human productions. Interestingly, the way that these models learn is also instructive. For instance, it has been noted that past tense learning follows a nonlinear, U-shaped, pattern. Specifically, children show a tendency to produce errors on irregular forms that they previously produced correctly,[27,28] and these errors often take the form of over-regularizations (e.g, **taked* instead of *took*). Proponents of a generative grammar perspective have suggested that this pattern occurs due to the overapplication of a rule to irregulars.[24,29] On this view, children initially use a memorization procedure to encode both present and past tense forms in their lexicon; later they discover the past tense rule but tend to overapply it to all forms; finally, they learn to use the rule only for regulars, and memorize only irregular past tenses.

Connectionist models provide a different way of conceptualizing this process. The RM86 model examined the effect of changing the size of the training vocabulary over time. Initially the model was trained on a small set of high-frequency verbs, many of them irregulars. While it showed good initial performance on these items, accuracy showed an initial decline after the training vocabulary size was subsequently increased. The reason for this was that the proportion of regular verbs in the model's vocabulary was initially quite small. Increasing the vocabulary size also increased the ratio of regular to irregular verbs, and consequently the model was able to pick up on the consistency of present–past mappings among these regulars. One interesting consequence is that the model tended to produce a high degree of over-regularization errors at this point in training, similar to what is observed in children. With further training, the over-regularization errors gradually disappeared, and the model regained high levels of accuracy on both regulars and irregulars.

Both the SM89 and RM86 models have raised a great deal of debate, much of it focusing on their ability to accurately capture human-like data, and the extent to which this reflects an overall failure of the architecture or inadequacies of details. As a result a number of follow-up models have been put forward aimed at accounting for a broader range of phenomena. Here we discuss a few of the key advances. The first of these is the use of improved phonological representations to encode word forms. Neither the RM86 nor the SM89 models generalized to nonwords as accurately as adult humans do. One reason appears to be the phonological coding scheme that was used. These models used individual units to encode triplets of phonemes in a word; for instance *sleep* is encoded by separate units that represent '#sl', 'sli', 'lip', and 'ip#'. More recent models have used a different approach in which a word's phonemes are divided in a more structured way into discrete consonant and vowel 'slots': for instance, Daugherty and Seidenberg[30] used a CCVVCC coding scheme that encodes *sleep* as [sli_p_], with underscores denoting unused slots. Using some variant of such a scheme yields much stronger generalization rates both in reading (e.g., Ref 31) and past tense (e.g., Ref 30).

Work growing out of the RM86 and SM89 models continues. For instance, the scope of these models has been expanded by including semantic

units that help to better account for lexical-level effects.[32,33] In addition, some researchers have advocated a 'neuroconstructivist' approach, which seeks to incorporate the interaction between the structure of the input and experience-dependent reorganization of the neural architecture within PDP models.[34] Importantly, improvements in how these models account for human data have not been accomplished at the cost of other desirable characteristics of the earlier models, which have generally been retained in the updated instantiations.

Also of note has been the extension of these models to languages other than English. With respect to morphology, this has included work on German plurals,[35] and noun marking in Serbian.[36] Such models address the extent to which learning is influenced by the structure of morphological systems. For instance, German includes many different types of irregulars, and the regular form appears to apply in a minority of cases. Likewise, Serbian marks nouns for number, gender, and case, but the system as a whole is only 'quasi-regular', such that no single form represents a classic regular. Instead, large patterns of similarity exist among forms, with many exceptional cases also integrated within these patterns. Both these statistical profiles can be accommodated by connectionist models, supporting the view that this approach is informative about a wide range of linguistic data.

Similarly, the connectionist approach has also begun to provide useful insights into reading in languages with different characteristics from English. For instance, Yang et al.[37] have adapted the SM89 architecture to reading in Chinese, where each symbol represents an entire word or concept. Given its dissimilarity to alphabetic languages, some have argued Chinese reading involves a solely lexical process in which words are recognized holistically without access to sublexical sound-based representations.[38] Nevertheless, closer inspection reveals the existence of phonological sub-units in Chinese that form semi-regular mappings between print and sound. And indeed, Yang et al.'s model does appear to take advantage of these somewhat hidden phonological regularities in Chinese. Their findings thus support the view that skilled reading of Chinese involves the same types of reading mechanisms as those used in English, and one does not need to assume different cognitive architectures to account for cross-linguistic data.

Neuropsychological Data in Children and Adults

There has also been a recent resurgence in interest in models of reading and past tense due to the observation of neuropsychological double dissociations in processing regular and irregular forms. The reading literature includes several classic descriptions of patients with acquired alexia following brain injury showing a specific deficit in reading either exception words (known as 'surface' dyslexia)[39] or nonwords (called 'deep' or 'phonological' dyslexia).[40] Likewise, developmental dyslexia in children has also been described as falling into surface/deep subtypes marked by difficulty reading exceptions or nonwords, respectively.[41]

One explanation of double dissociations assumes functional modularity, in which separate neurocognitive systems are responsible for processing rules and exceptions, and which are differentially impaired in different syndromes. Indeed, such findings may at first appear inconsistent with a connectionist view, in which a single mechanism is used to encode all forms. To address this, contemporary connectionist approaches partially concede a certain degree of specificity, while still treating all types of items in a uniform architecture in which damage to different parts of the network can differentially affect items of different types. Plaut et al.[31] revisited the SM89 reading model (Figure 5) by implementing it with completely interconnected orthographic, phonological and semantic units. They showed that different types of reading difficulties arise as a result of damage to each of these groups of units, or by severing different sets of connections among them. For example, damage to connections linking orthography to phonology yielded specific difficulty with irregulars, simulating surface dyslexia. In contrast, damage to the phonological layer yielded a distinct deficit in reading nonwords, simulating phonological dyslexia. Similarly Harm and Seidenberg[42] took a similar approach to simulating developmental dyslexia by implementing different types of pre-existing damage to a connectionist model of reading prior to learning.

Neuropsychological dissociations have also been observed in past tense processing. For instance, Ullman et al.[43] identified patients who had difficulty producing nonword forms (which typically take a regular -ed ending), and others that had specific difficulty with irregular forms. This again could suggest damage to dissociable brain mechanisms subserving rules and exceptions, respectively. Nonetheless these findings also do not preclude a connectionist explanation. Building on the earlier work with reading models, Joanisse and Seidenberg[33] suggested dissociations in past tense arise from damage to brain regions responsible for phonology or semantics. They proposed a model of past tense that included both a phonological component somewhat similar to the earlier SM89 model, and a semantics layer used to

uniquely encode meanings of individual word forms. Different types of brain damage were simulated by artificially lesioning groups of units responsible for coding either phonological or semantic knowledge. As predicted, the two types of damage yielded distinct patterns of impairment on nonwords and irregulars, due to differences in the degree to which these types of forms rely on phonological and semantic information.

The explanation for why irregulars and nonwords can be differentially impaired in these simulations is as follows: although connectionist networks are homogeneous in their use of a connectionist mechanism, they do encode different types of information across different sets of units and connections. Different forms rely to differing degrees on these types of knowledge, and consequently can be impacted to greater or lesser degrees by damage to a specific component of the model. Specifically, nonwords rely more strongly on phonological knowledge due to the importance of phonology in learning spelling-to-sound consistency. In contrast, irregulars have an inconsistent phonological relationship between present and past tense forms and thus are less susceptible to phonological damage. Instead, the network relies on support from other mechanisms (e.g., the semantic layer in the Joanisse and Seidenberg[33] past tense model) to learn the idiosyncrasies of irregular forms' spelling-to-sound mappings.

Several researchers have noted that connectionist models also provide a useful basis for understanding developmental deficits in forming past tenses, other word inflections, or in learning to read. One approach[44,45] focuses on the effects of phonological deficits that might make it difficult to detect or represent aspects of speech phonology, thereby impairing access to the subtle phonetic cues used to mark regularly inflected forms in English. The past tense marker—often a subtle 't' or 'd' sound added to the end of the base wordform is particularly weak phonetically, and this may contribute to difficulty mastering the regular pattern. Typically, exceptions differ more from their regular counterparts, due in many cases to a vowel change, possibly along with other changes (e.g., *see–saw* and *buy–bought*) and would thus be less susceptible to perceptual or phonological difficulties. Another approach focuses on network characteristics that can differentially impact exceptional and regular forms in single-system models like the RM86 network. Indeed, Thomas and Karmiloff-Smith[46] have argued that double dissociations may reflect distinct anomalous distortions of a single underlying network, differentially affecting regular and exception forms, rather than separate mechanisms for regular forms and exceptions.

CONNECTIONIST APPROACHES TO SYNTACTIC AND SEMANTIC PROCESSING

In addition to addressing the processing of single words, connectionist approaches have also been extended to examine syntactic and semantic processing. One key theme of early work[47,48] was the demonstration that fairly simple connectionist models could learn to rely on long-distance syntactic dependencies, such as number agreement between the head noun and main-clause verb in a sentence like 'The boys who saw the girl like ice-cream' (note that 'like' must agree with 'boys' though this noun is further away from the verb than is 'girl'). Such phenomena had long been held to demand a domain-specific language acquisition device preprogrammed with knowledge of core principles of language. The use of a simple connectionist model that simply learns to predict successive elements in word sequences to capture such dependencies was a dramatic departure from this thinking, and suggested that no such device was really necessary.

Related research has built on these early successes, capturing a wide range of phenomena in on-line language processing, including the role of word meaning as well as syntactic information in correctly uncovering the role of relationships among the constituents of sentences (e.g., Ref 49). As one example of such a phenomenon, consider the sentence 'The spy shot the policeman with the _____'. If the final word is 'revolver', readers interpret this item as the instrument used to carry out the action (shooting). If the final word is 'binoculars', however, they interpret this item as an object associated with (perhaps being held by) the policeman. The assignments reverse if 'shot' is replaced by 'saw'. Many studies show that humans are highly sensitive to these aspects of word meaning, using them on-line to affect their interpretations of such sentences. Furthermore, differences in the frequency with which nouns enter into particular roles with respect to particular verbs affect the speed of sentence comprehension and the likelihood of temporary misinterpretation.[49] To date, most of these models have focused on processing of sentences in normal adults, and full accounts of disorders of sentence processing remain to be developed within a connectionist framework. That said, there are a number of connectionist models addressing both receptive and productive aspects of the deficits exhibited by patients suffering from aphasia.[50,51]

Another body of connectionist work focuses on disorders of semantic processing in patients with a condition often called 'semantic dementia' (SD). This condition can arise from any one of several different

progressive neuropathological disorders, when these affect a particular set of brain areas centered around the anterior inferior temporal cortex.[52] The disorder appears to affect the very knowledge of the things words refer to, and this might lead some to think such a deficit should be excluded from a discussion of perspectives on and models of language. However, as one would expect from the interactive perspective inherent in the connectionist approach, this object knowledge impairment leads to quite a striking pattern of language impairment.

Along with a progressive loss of object knowledge, first affecting infrequent and atypical things, SD patients also show a striking preference for typicality, both in words and objects.[53] Given a choice between 'frute' or 'fruit' in a lexical decision task, SD patients tend to choose 'frute'—the item with the more typical spelling. Similarly, given a choice between a typicalized elephant (one whose ear has been replaced by the more typical ear of a monkey) and a real elephant in an object decision task, they tend to choose the typicalized elephant. Parallel preferences for typical and linguistically regular items occur in single word reading, past tense inflection, and other tasks. As we should expect based on the properties of the connectionist models discussed throughout this article, corresponding deficits all arise from damage to units or connections in a simple interactive connectionist network that learns from paired presentations of visual and nonvisual semantic information about objects and phonological and orthographic information about these objects' names.[54,55]

DEEP LEARNING

One of the most exciting recent developments in connectionism comes from the applied field in which 'deep networks' are being used to classify large and complex datasets.[56] These models involve multiple hidden layers that mediate the input from the output. They are trained using backpropagation and related algorithms, in a way that allows them to develop increasingly abstract representations of the data and thus discover complex and nonobvious patterns within datasets. The successes of these networks are typically discussed in terms of machine learning applications, as in voice recognition[57] and visual object categorization.[58] However, these models are also being successfully applied to issues in sentence processing, including parsing and interpretation of the sentiment expressed in a sentence.[59,60] Thus, there is the strong potential to apply deep learning mechanisms to questions in psycholinguistics, and we would expect that the results will provide useful insights into the way in which language is learned, represented and processed.

One challenge for this view is how we might analyze the organization of deep learning networks. They consist of very large sets of artificial neurons and their organization into multiple hidden layers might also add to their complexity. As a result, one might suppose that understanding how and why they produce certain behaviors could be especially difficult if not impossible. That said, the analytic tools at our disposal are also continuing to develop, and there have already been some proposals of ways in which we can understand the performance of these systems.[61] Indeed, the concern about understanding and analyzing complex connectionist networks is one that has been raised even from the outset. However the problem has not proven to be insurmountable so far, and there is every reason to think that we will continue to develop appropriate analytic approaches as these networks continue to scale up.

OTHER STATISTICAL APPROACHES

The connectionist enterprise represents a shift from symbol-based accounts of the mind to a more probabilistic or statistical approach. However, this is not the only approach that takes such a view. For instance, the Bayesian view of cognition also seeks to account for a range of language and cognition behaviors via sets of probabilistically weighted constructs.[62] In some ways, the Bayesian and connectionist approaches appear compatible, given their commitment to the idea that behavior can be explained via the interaction of multiple sources of probabilistically weighted information.

That said, there are differences between connectionist approaches and some Bayesian models. For example, many Bayesian models build in one or a subset of specifically structured 'hypothesis spaces' while connectionist models can instead be seen as exploring a more continuous hypothesis space that can capture a wider range of representational possibilities and can more adequately capture the fact that natural data only approximates any specific structure type. In addition, Bayesian models often rely on sources of information that may be abstractions far removed from the actual mechanisms that serve to explain patterns of learning and behavior. For instance, noting that the token frequency of a word is an important predictor of reading times in a Bayesian model has, in our view, much less explanatory value than showing the learning mechanism in a PDP model that explains precisely why these frequency effects emerge. For further discussion of this issue, see discussion in Jones and Love,[63] and McClelland et al.[64]

CONCLUSIONS

In this review, we have presented an overview of connectionist modeling of language, focusing both on early efforts and more recent developments. The approach provides a distinctive perspective on language learning and representation—one that is relevant not only to language processing as it occurs in typical adults but also to acquisition and to disorders of language processing. Specifically, it eschews ideas of domain-specific representational and learning mechanisms, and suggests instead that we can understand language phenomena using a simple set of cognitive principles. The approach also seeks to put the 'learning' back into language learning phenomena by providing a mechanism that discovers complex linguistic patterns thanks to statistically structured patterns in the input. The approach explains the 'shape of change'—nonlinearities observed in development emerge naturally out of the way in which these types of models learn information—as well as deficits that can arise from effects of damage after learning or anomalies in the characteristics of the developing system. Connectionist models also address a wide range of phenomena in sentence processing, and have proven especially useful in modeling semantic learning and disorders of knowledge representations in SD.

A distinct challenge of connectionism is the concern that one must become a modeler in order to incorporate into one's research program the theoretical assumptions of the connectionist enterprise. It is becoming increasingly clear that this is not the case. Instead, many new behavioral studies of language learning, processing and impairment have begun to incorporate many of the principles of connectionism. Especially noteworthy is the increasing interest in studies of how statistical learning allows infants to rapidly acquire phonological categories,[65] learn to segment words from the continuous speech stream[66] and to analyze the distributional properties of these words to acquire syntactic representations.[67]

NOTES

[a] Although we would argue these are fundamental guiding principles of the connectionist enterprise, there is variability in the extent to which individual models reflect each of them. For instance with respect to assumption 1, some models do use localist representations in which single units represent whole concepts or categories rather than the distributed representations in which items are represented by ensembles of units that also participate in representing other items. In some cases, these are simplifying assumptions made in the interest of keeping models computationally tractable. In other cases,[6] this localist coding scheme reflects a strong theoretical claim about the nature of mental representations.

[b] Note that McClelland and Elman proposed two implementations of TRACE to account for somewhat different phenomena. Here we focus on the implementation that was named 'TRACE-II' in the original work.

ACKNOWLEDGMENTS

MFJ is supported by operating grants from the Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council (Canada).

REFERENCES

1. Chomsky N. On certain formal properties of grammars. *Inf control* 1959, 2:137–167.
2. Newmeyer FJ. *Language Form and Language Function*. Cambridge, MA: MIT Press; 1998.
3. Pinker S, Jackendoff R. The faculty of language: What's special about it? *Cognition* 2005, 95:201–236.
4. McClelland JL, Rumelhart DE, The PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. II. Cambridge, MA: MIT Press; 1986.
5. Smolensky P. Grammar-based connectionist approaches to language. *Cognit Sci* 1999, 23:589–613.
6. Bowers JS, Vankov II, Damian MF, Davis CJ. Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychol Rev* 2014, 121:248–261.
7. Piatelli-Palmarini M. *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. London: Routledge & Kegan Paul; 1980.
8. Forster KI. Accessing the mental lexicon. In: Wales RJ, Walker E, eds. *New Approaches to Language Mechanisms*. Amsterdam: North-Holland; 1976, 257–287.
9. Marslen-Wilson W. Functional parallelism in spoken word recognition. *Cognition* 1987, 25:71–102.

10. McClelland JL, Elman JL. The TRACE model of speech perception. *Cogn Psychol* 1986, 18:1–86.
11. Warren RM. Perceptual restoration of missing speech sounds. *Science* 1970, 167:392–393.
12. Ganong WF. Phonetic categorization in auditory perception. *J Exp Psychol Hum Percept Perform* 1980, 6:110–125.
13. Allopenna PD, Magnuson JS, Tanenhaus MK. Tracking the time course of spoken word recognition: evidence for continuous mapping models. *J Mem Lang* 1998, 38:419–439.
14. Dahan D, Magnuson JS, Tanenhaus MK. Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cogn Psychol* 2001, 42:317–367.
15. Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. *Science* 1995, 268:1632–1634.
16. Tanenhaus MK, Magnuson JS, Dahan D, Chambers C. Eye movements and lexical access in spoken-language comprehension: evaluating a linking hypothesis between fixations and linguistic processing. *J Psycholinguist Res* 2000, 29:557–580.
17. Magnuson JS, Tanenhaus MK, Aslin RN, Dahan D. The microstructure of spoken word recognition: studies with artificial lexicons. *J Exp Psychol Gen* 2003, 132:202–227.
18. McClelland JL, Mirman D, Holt LL. Are there interactive processes in speech perception? *Trends Cogn Sci* 2006, 10:363–369.
19. Grosjean F. Exploring the recognition of guest words in bilingual speech. *Lang Cogn Proc* 1988, 3:233–274.
20. Dijkstra A, Van Heuven WJB. The BIA model and bilingual word recognition. In: Grainger J, Jacobs A, eds. *Localist Connectionist Approaches to Human Cognition*. Hillsdale, NJ: Lawrence Erlbaum; 1998, 189–225.
21. Beauvillain C, Grainger J. Accessing interlexical homographs: some limitations of a language-selective access. *J Mem Lang* 1987, 26:658–672.
22. Strauss TJ, Harris HD, Magnuson JS. jTRACE: a reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behav Res Methods* 2007, 39:19–30.
23. Hannagan T, Magnuson JS, Grainger J. Spoken word recognition without a TRACE. *Front Psychol* 2013, 4:563. doi:10.3389/fpsyg.2013.00563.
24. Pinker S. Rules of language. *Science* 1991, 253: 530–535.
25. Rumelhart DE, McClelland JL. On learning the past tenses of English verbs. In: McClelland JL, Rumelhart DE, The PDP Research Group, eds. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. II Chapter 18. Cambridge, MA: MIT Press; 1986, 216–271.
26. Seidenberg MS, McClelland JL. A distributed, developmental model of word recognition and naming. *Psychol Rev* 1989, 96:523–568.
27. Bybee JL, Slobin DI. Rules and schemas in the development and use of the English past tense. *Language* 1982, 58:265–289.
28. Kuczaj S. The acquisition of regular and irregular past tense forms. *J Verbal Learning Verbal Behav* 1977, 16:589–600.
29. Marcus GF, Pinker S, Ullman M, Hollander M, Rosen TJ, Xu F, Clahsen H. Overregularization in language acquisition. *Monogr Soc Res Child Dev* 1992, 57:1–178.
30. Daugherty K, Seidenberg MS. Rules or connections? The past tense revisited. In: *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum; 1992, 259–264.
31. Plaut DC, McClelland JL, Seidenberg M, Patterson KE. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol Rev* 1996, 103:56–115.
32. Harm MW, Seidenberg MS. Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychol Rev* 2004, 111:662–720.
33. Joanisse MF, Seidenberg MS. Impairments in verb morphology after brain injury: a connectionist model. *Proc Natl Acad Sci USA* 1999, 96:7592–7597.
34. Westermann G, Ruh N. A neuroconstructivist model of past tense development and processing. *Psychol Rev* 2012, 119:649–667.
35. Hahn U, Nakisa RC. German inflection: Single route or dual route? *Cogn Psychol* 2000, 41:313–360.
36. Mirkovic J, Seidenberg MS, Joanisse MF. Probabilistic nature of inflectional structure: insights from a highly inflected language. *Cognit Sci* 2011, 35:638–681.
37. Yang J, McCandliss BD, Shu H, Zevin JD. Simulating language-specific and language-general effects in a statistical learning model of Chinese reading. *J Mem Lang* 2009, 61:238–257.
38. Zhou X, Marslen-Wilson W. The nature of sublexical processing in reading Chinese characters. *J Exp Psychol Learn Mem Cogn* 1999, 25:819–837.
39. Patterson KE, Marshall JC, Coltheart M. *Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading*. London: Lawrence Erlbaum; 1985.
40. Beauvois MF, Derouesné J. Phonological alexia: Three dissociations. *J Neurol Neurosurg Psychiatry* 1979, 42:1115–1124.
41. Manis F, Seidenberg M, Doi L, McBride-Chang C, Peterson A. On the basis of two subtypes of developmental dyslexia. *Cognition* 1996, 58:157–195.

42. Harm MW, Seidenberg MS. Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychol Rev* 1999, 106:491–528.
43. Ullman M, Corkin S, Coppola M, Hickok G, Growdon JH, Koroshetz WJ, Pinker S. A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *J Cogn Neurosci* 1997, 9:289–299.
44. Leonard L. *Children with Specific Language Impairment*. Cambridge, MA: MIT Press; 1998.
45. Tallal P, Miller S, Fitch R. Neurobiological basis of speech: a case for the preeminence of temporal processing. In: Tallal P, Galaburda AM, Llinas RR, von Euler C, eds. *Temporal Information Processing in the Nervous System: Special Reference to Dyslexia and Dysphasia*. New York, NY: New York Academy of Sciences; 1993, 27–47.
46. Thomas MSC, Karmiloff-Smith A. Modelling language acquisition in atypical phenotypes. *Psychol Rev* 2003, 110:647–682.
47. Elman JL. Finding structure in time. *Cognit Sci* 1990, 14:179–211.
48. Elman JL. Distributed representations, simple recurrent networks, and grammatical structure. *Mach Learn* 1991, 7:195–224.
49. MacDonald MC, Pearlmutter NJ, Seidenberg MS. The lexical nature of syntactic ambiguity resolution. *Psychol Rev* 1994, 101:676–703.
50. Dell GS, Schwartz MF, Martin N, Saffran EM, Gagnon DA. Lexical access in aphasic and nonaphasic speakers. *Psychol Rev* 1997, 104:801–838.
51. Gotts SJ, Plaut DC. Connectionist approaches to understanding aphasic perserveration. *Semin Speech Lang* 2004, 25:323–334.
52. Pereira JMS, Williams GB, Acosta-Cabronero J, Pengas G, Spillantini MG, Xuereb JH, Hodges JR, Nestor PJ. Atrophy patterns in histologic vs. clinical groupings of frontotemporal lobar degeneration. *Neurology* 2009, 72:1653–1660.
53. Rogers TT, Lambon Ralph MA, Hodges J, Patterson K. Object recognition under semantic impairment: the effects of conceptual regularities on perceptual decisions. *Lang Cogn Proc* 2003, 18:625–662.
54. Dilkina K, McClelland JL, Plaut DC. A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cogn Neuropsychol* 2008, 25:136–164.
55. Rogers TT, Lambon Ralph MA, Garrard P, Bozeat S, McClelland JL, Hodges JR, Patterson K. The structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychol Rev* 2004, 111:205–235.
56. Hinton G, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput* 2006, 18:1527–1554.
57. Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition. *Science* 2009, 4:1–9. doi:10.4249/scholarpedia.5947.
58. Le QV, Ranzato MA, Monga R, Devin M, Chen K, Corrado GS, Dean J, Ng AY. Building high-level features using large scale unsupervised learning. In: *Proceedings of the International Conference on Machine Learning*, Edinburgh, Scotland, June 26–July 1, 2012.
59. Socher R, Bauer J, Manning CD, Ng AY. Parsing with compositional vector grammars. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria, August, 2013, 455–465. Association for Computational Linguistics. Available at: http://www.aclweb.org/anthology/P13-1045. (Accessed December 1, 2014).
60. Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, October, 2013, 1631–1642. Association for Computational Linguistics. Available at: http://www.aclweb.org/anthology/D13-1170. (Accessed December 1, 2014).
61. Saxe AM, McClelland JL, Ganguli S. Learning hierarchical category structure in deep neural networks. In: Knauff M, Paulen M, Sebanz N, Wachsmuth I, eds. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society; 2013, 1271–1276.
62. Chater N, Manning CD. Probabilistic models of language processing and acquisition. *Trends Cogn Sci* 2006, 10:335–344.
63. Jones M, Love BC. Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav Brain Sci* 2011, 34:169–231.
64. McClelland JL, Botvinick MM, Noelle DC, Plaut DC, Rogers TT, Seidenberg MS, Smith LB. Letting structure emerge: connectionist and dynamical systems approaches to understanding cognition. *Trends Cogn Sci* 2010, 14:348–356.
65. Maye J, Werker JF, Gerken L. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 2002, 82:B101–B111.
66. Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. *Science* 1996, 274:1296–1298.
67. Chemla E, Mintz TH, Bernal S, Christophe A. Categorizing words using frequent frames: what cross-linguistic analyses reveal about distributional acquisition strategies. *Dev Sci* 2009, 12:396–406.