Connectionist Models of Language Processing

Douglas L. T. Rohde
Massachusetts Institute of Technology
David C. Plaut
Carnegie Mellon University and the Center for the Neural Basis of Cognition

March 14, 2003

Abstract

Traditional approaches to language processing have been based on explicit, discrete representations which are difficult to learn from a reasonable linguistic environment—hence, it has come to be accepted that much of our linguistic representations and knowledge is innate. With its focus on learning based upon graded, malleable, distributed representations, connectionsit modeling has reopened the question of what could be learned from the environment in the absence of detailed innate knowledge. This paper provides an overview of connectionist models of language processing, at both the lexical and sentence levels.

1 Introduction

Although connectionist models have been applied to the full range of perceptual, cognitive, and motor domains (see McClelland, Rumelhart, & PDP Research Group, 1986; Quinlan, 1991; McLeod, Plunkett, & Rolls, 1998), it is in their application to language that they have evoked the most interest and controversy (e.g., Pinker & Mehler, 1988). This is perhaps not surprising in light of the special role that language plays in human cognition and culture. It also stems in part from the considerable difference in goals and methods between linguistic and psychological approaches to the study of language. This rift goes deeper than a simple dichotomy of emphasizing competence versus performance (Chomsky, 1957)—it cuts to the heart of the question of what it means to know and use a language (Seidenberg, 1997).

Traditional approaches to language processing have been based on explicit, discrete representations which are difficult or impossible to learn from a reasonable linguistic environment (Gold, 1967). Therefore, it has come to be accepted that much of our linguistic representations and knowledge is innate. With its focus on learning based upon graded, malleable, distributed representations, connectionism has reopened the question of what could be learned from the environment in the absence of detailed innate knowledge. Although the need to learn internal representations potentially gives connectionist networks great power and flexibility, it also introduces limitations. These limitations are important and, ideally, will reflect limitations observed in human language processing.

From a connectionist perspective, performance is not an imperfect reflection of some abstract competence, but rather the behavioral manifestation of the internal representations and processes of actual language users: Language is as language does. In this regard, errors in performance (e.g., "slips of the tongue"; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997) are no less valid than skilled language use as a measure of the underlying nature of language processing. The goal is not to abstract away from performance but to articulate computational principles that account for it.

A major attraction of the connectionist approach to language, apart from its natural relation to neural computation, is that the very same processing mechanisms apply across the full range of linguistic structure. This paper provides an overview of connectionist models of language processing, at both the lexical and sentence levels.

2 Lexical Processing

2.1 Phonological development

Although the use of language seems straightforward to adult native speakers, an infant must solve numerous difficult computational problems in learning to understand and produce speech, stemming from the fact that speech is extended in time, highly variable and, at a morphemic level, has no systematic relation to its underlying meaning. Moreover, infants must learn to produce comprehensible speech without any direct articulatory instruction or feedback.

Plaut and Kello (1999) proposed a framework for

[*] Draft Version. The final version appeared in Takahashi Tsuzuki (Ed.), Special issue of *Cognitive Studies, Japan*, 10(1), 10-28.

phonological development in which phonology mediates among acoustic, articulatory, and semantic representations in the service of both comprehension and production. A critical aspect of the approach is that, given the absence of direct articulatory feedback, learning to produce speech is driven by indirect feedback derived from the comprehension system—that is, from the acoustic, phonological, and semantic consequences of the system's own articulations (Locke, 1983; Menn & Stoel-Gammon, 1995; Studdert-Kennedy, 1993). This is accomplished by learning an internal *forward model* of the physical processes that relate articulation to acoustics (Jordan & Rumelhart, 1992). Such a model is learned by executing a variety of articulations, predicting how they will sound, and then adapting the model based on the discrepancy between this prediction and the actual resulting acoustics. In the infant, the forward model is assumed to develop primarily as a result of reduplicated and variegated babbling in the second half of the first year (Vihman, 1996). Once developed, the forward model can be used to convert acoustic feedback (i.e, whether an utterance sounded right) into the articulatory feedback necessary to train speech production (Perkell, Matthies, Svirsky, & Jordan, 1995). An implementation of the framework, in the form of a simple recurrent network (Elman, 1991a), learned to comprehend, imitate, and intentionally name a corpus of 400 monosyllabic words, and its speech errors in development were similar to those of young children.

2.2 Morphology

Most linguistic domains are *quasi-regular* in that there is considerable systematicity between inputs and outputs but also numerous exceptions. A standard assumption is that systematic linguistic knowledge takes the form of explicit rules and that items which violate the rules are handled by a separate associative mechanism (see Pinker, 1999). Connectionist modeling provides an alternative view, in which all items coexist within a single system whose representations and processing reflect the relative degree of *consistency* in the mappings for different items.

A key battleground in the debate between these two views of the language system has been the relatively constrained domain of English inflectional morphology—specifically, forming the past-tense of verbs. Rumelhart and McClelland (1986) attempted to reformulate the issue away from a sharp dichotomy between explicit rules (add –ed; e.g., WALK/WALKED) and exceptions (e.g., SING/SANG, DRINK/DRANK, GO/WENT), and toward a view that emphasizes the graded structure relating verbs and their inflections. They developed a connectionist model that learned a direct association between the phonology of all types of verb stems and the phonology of their past-tense forms. Although this initial model had numerous limitations (Pinker & Prince, 1988), many of these have been addressed in subsequent simulation work (Cottrell & Plunkett, 1995; MacWhinney & Leinbach, 1991; Marchman, 1993; Plunkett & Marchman, 1991, 1993, 1996). Moreover, applications of connectionist models to aspects of language disorders (Joanisse & Seidenberg, 1999; Hoeffner & McClelland, 1993; Marchman, 1993) and language change (Hare & Elman, 1995) demonstrate the ongoing extension of the approach to account for a wider range of language phenomena.

Derivational morphology has also been a context in which connectionist models have contrasted with more symbolic, rule-based accounts. On a distributed connectionist approach, derivational morphology reflects a learned sensitivity to the systematic relationships among the surface forms of words and their meanings. Consistent with this perspective, Gonnerman, Andersen, and Seidenberg (submitted; Seidenberg & Gonnerman, 2000) have demonstrated graded effects of both semantic and formal similarity in cross-modal morphological priming. By the same token, however, findings of non-semantic morphological priming in morphologically rich languages like Hebrew (e.g., Frost, Deutsch, & Forster, 2000) are typically interpreted as being problematic for the connectionist account. To evaluate whether this interpretation is valid, Plaut and Gonnerman (2000) carried out simulations in which a set of morphologically related words varying in semantic transparency were embedded in either a morphologically rich or impoverished artificial language. They found that morphological priming increased with degree of semantic transparency in both languages. Critically, priming extended to semantically opaque items in the morphologically rich language (consistent with findings in Hebrew) but not in the impoverished language (consistent with findings in English). Such priming arises because the processing of all items, including opaque forms, is influenced by the degree of morphological organization of the entire system. These findings suggest that, rather than being challenged by the occurrence of non-semantic morphological effects in morphologically rich languages, the connectionist approach may provide an explanation for the cross-linguistic differences in the occurrence of these effects.

2.3 Word reading

Many of the issues concerning quasi-regularity in morphological processing also arise in the context of word reading. As in morphology, the spelling-sound correspondences of English are highly systematic but admit many exceptions (e.g., HAVE, PINT, YACHT) and, as in morphology, researchers have proposed separate mechanisms for processing regular and exception items (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001).

Plaut, Seidenberg, McClelland, and Patterson (1996) developed a series of connectionist simulations in support of an alternative conception of language knowledge and processing in which all items coexist within a single system whose representations and processing reflect the relative degree of *consistency* in the mappings for different items. Different types of information about a word—orthographic, phonological, and semantic—are represented as distributed patterns over separate groups of units. In performing a task like reading aloud, orthographic information influences phonological output via two pathways: a *phonological* pathway that maps orthography to phonology directly (via hidden units), and a *semantic* pathway that maps first to meaning and then to phonology.

An early connectionist simulation of the phonological pathway (Seidenberg & McClelland, 1989) provided a good account of word reading but was poor at pronouncing word-like nonwords (e.g., MAVE). Plaut et al. (1996) showed that the limitations of this preliminary model stemmed not from any general failing of connectionist networks in quasi-regular domains, but from its use of poorly structured orthographic and phonological representations. When representations were used that condense the regularities between orthography and phonology by incorporating graphotactic and phonotactic constraints, networks were able to learn to pronounce both regular and exception words, and yet also pronounce nonwords as well as skilled readers.

Although implementations of the phonological pathway alone can learn to pronounce words and nonwords effectively, skilled reading requires the combined support of both the semantic and phonological pathways. This consideration has important implications for understanding acquired surface dyslexia, which typically arises from a semantic impairment. Surface dyslexic patients often misread low-frequency exception words by producing a more "regular" pronunciation (e.g., reading SEW as "sue"). Plaut et al. (1996) demonstrated that that surface dyslexia can arise as a result of the natural limitations of an intact phonological pathway that had learned to rely on semantic support, when semantics is impaired by brain damage. In closely related work, Harm and Seidenberg (2001) demonstrated how the complementary disorder of *phonological* dyslexia—in which nonword reading is impaired relative to word reading—can arise as a result of phonological rather than semantic damage.

Finally, Harm and Seidenberg (1999) showed how the same framework can account for both normal and disordered reading acquisition. Four issues were examined: the acquisition of phonological knowledge prior to reading, how this knowledge facilitates learning to read, phonological and non-phonological bases of dyslexia, and effects of literacy on phonological representation. Compared with simple feedforward networks, representing phonological knowledge in an attractor network yielded improved learning and generalization. Phonological and surface forms of developmental dyslexia, which are usually attributed to impairments in distinct lexical and nonlexical processing "routes," were derived from different types of damage to the network.

In summary, connectionist models of lexical processing have demonstrated that phenomena that appear to require explicit, stipulated representations, or multiple processing mechanisms, can instead by captured in a natural fashion by the basic computational properties of distributed networks learning in quasi-regular domains. Indeed, many of these properties also provide leverage in understanding language performance at the sentence level.

3 Sentence Processing

Most sentence processing models have been designed to address one of four major language tasks: parsing, word prediction, comprehension, or production. The models have been organized here by the primary task for which they were designed, rather than in chronological order.

3.1 Parsing

Parsing, or producing a syntactic, structural description of a sentence from its surface form, is the one sentence processing task that has received the most attention from the symbolic community. Thus, it should not be surprising that many of the connectionist parsing systems found in the literature are essentially symbolic models implemented transparently in connectionist hardware. Learning has not played a major role in most of these parsing models for two main reasons. First, most connectionist parsing models have been localist, meaning that each unit represents an explicit state or bit of information. This architecture lends itself to hand-designed weight structures but not to the easy design of effective learning environments. More critically, teaching a model to produce an explicit parse of a sentence requires, for most systems, training data labeled with correct parsing information. Few believe that such information is actually available to the child, so models which rely on it are of questionable relevance to human learning.

3.1.1 Localist parsing models

Unlike some of the later, more transparently symbolic approaches, the first significant proposal for a connectionist model of parsing, Small, Cottrell, and Shastri (1982), which was based on the ideas of McClelland and Rumelhart (1981), stresses the importance of interaction between syntactic information and semantics over more standard parsing theories that stress compartmentalism and serial processing (Frazier, 1979; Fodor, 1983). This model, implemented in later work (Cottrell, 1985b), is not actually a full parser but is designed for word-sense disambiguation. It uses localist units to represent lexical items, individual word senses, and case roles, and these units excite or inhibit one another through a set of hand-designed connections. Because of this, the model is not easily expandable to larger vocabularies or complex linguistic structures.

Cottrell (1985a) extended the earlier work with the addition of a full-fledged syntactic parsing network, which can be generated automatically given a grammar. Concepts are associated with case roles by means of localist *binder* units. There is a unit for each concept/role pair and these units mutually inhibit one another. Units in the syntactic portion of the network represent the non-terminal symbols of the context-free grammar, and their interconnections reflect the possible productions in the grammar. The model is interesting in that it is able to process sentences presented in a temporal sequence and makes use of interacting top-down and bottom-up information. However, it has a number of limitations. As is a common problem with other models that make use of case-roles, the model does not appear capable of handling sentences with multiple verbs. It can also handle only fixed-length sentences and requires constituent recognizers with duplicated and possibly non-connectionist control structures.

Several other localist parsing models were produced at the same time, and are primarily instantiations of symbolic parsers in the hardware of simple, interacting units. All of these models have difficulty handling recursive, context-free structure and the potentially long sentences that can result from it. One solution, adopted by Waltz and Pollack (1985) and Howells (1988), is to produce a customized network architecture on the basis of the grammar and the actual sentence being parsed. This network then settles into a structural interpretation of the sentence. Although the implemented model was purely localist, Waltz and Pollack proposed that concepts should not be represented by single nodes but by distributed patterns of “microfeatures,” a suggestion that would be adopted in later connectionist modeling. The model of Nakagawa and Mori (1988) also involved generation of the network on-the-fly, but it did so in the course of parsing, essentially implementing a left-corner parser. Although on-the-fly generation of networks is a neat trick, this seems rather implausible as a model of the human parsing mechanism.

An alternative approach to parsing in localist networks is to construct a single large network that is able to parse all sentences, but only up to a fixed length. Aiming to produce a network that is deterministic, fast, and guaranteed to work, Fanty devised a way to implement the CYK dynamic-programming, context-free parsing algorithm (Younger, 1967) in a localist network. It essentially contains a unit for every pairing of a non-terminal with a sub-sequence of the sentence. The network operates in two passes: a bottom-up phase in which units for increasingly longer sub-sequences become active if their non-terminal could have produced that sequence and a top-down phase in which units that do not fit within a coherent parse are silenced. Aside from the fact that it has an upper bound on sentence length, a major drawback of this model is that it is unable to resolve global parsing ambiguities. Another questionable property is its extensive redundancy. A model capable of parsing sentences up to length N would require roughly $N^2/2$ units for every non-terminal in the grammar. That is, there is a different NP unit for every place an NP might occur in a sentence. Although the model is not able to learn entire grammars, Fanty discussed how small errors could be corrected through learning and Rager (1992) described a localist model based on Fanty’s but designed to handle “extragrammatical,” or slightly incorrect, sentences.

Selman and Hirst (1985, 1994) presented a model that differs from other early connectionist parsers in that it uses a variation on the Boltzmann machine (Fahlman, Hinton, & Sejnowski, 1983). The rules of a context-free grammar are implemented in the network by means of syntactic binder units that inhibit one another and excite other units representing symbols that participate together in a production. The use of simulated annealing, while very slow, allows the network to gradually settle into the correct parse with high probability. But this model, too, requires sentences to be bounded in length and it relies on redundant structure. Due to the proliferation of binder units, the size of the network may grow intractably with more complex grammars. Furthermore, although the authors suggested it as a next step, this model does not incorporate semantic information and it is not clear how it would deal with syntactic ambiguity.

Charniak and Santos (1987) described another localist parsing model that differs from the others in its use of a sliding input window. This allows the network theoretically to handle sentences of unbounded length but hinders the ability of the model to process long-distance dependencies, such as those surrounding center-embeddings. Although the model was successfully implemented for a very simple grammar, it is not clear that its parsing heuristics would be sufficient to handle more complex grammars. The model also uses parts of speech rather than lexical inputs and is thus clearly unable to incorporate semantics or resolve syntactic ambiguities.

Because they lack the representational capacity of either a true symbolic system or a network with distributed representations, localist networks are severely impaired as language processors. The standard solution is to place a hard limit on sentence length and rely on extensive redundant structure, and an alternative is to generate the network on the basis of sentence length, which circumvents the problem of length but renders the model implausible biologically. Localist models are also heavily dependent on human design or predefined grammars, with little capability to learn. With a few exceptions, these models do not process sentences sequentially, as humans seem to do. Furthermore, they do not easily permit the inclusion of semantic or contextual constraints which are important sources of information in parsing ambiguous sentences (McClelland, St. John, & Taraban, 1989).

3.1.2 Hybrid and distributed parsing models

While the last decade has seen a number of hybrid connectionist/symbolist parsing models, only a few will be mentioned here. One trend in such models is the replacement of sub-components within a modular symbolic system with trained networks. The CDP model of Kwasny and Faisal (1990) is a modification of the PARSIFAL deterministic parser (Marcus, 1980). Several of the components of this rule-based parser were removed and replaced with a connectionist network, which was trained to suggest actions to be taken by the symbolic components of the model given the parsing context. Although the model was reportedly able to process ungrammatical and lexically ambiguous sentences in an appropriate way, it is not clear what effect the network component played in the behavior of the model. A similar approach was taken by Tepper, Powell, and Palmer-Brown (2001) in designing a shift-reduce parser with connectionist modules but symbolic control.

Wermter and Weber (1994, 1997) and Weber and Wermter (1996) were interested in creating a system that was robust to extragrammatical sentences. Their SCREEN model is a complex, highly modular, system with most of the modules consisting of trained networks. Rather than producing full parse trees, the SCREEN model generates a *flat* syntactic and semantic parse. The model was trained and tested on spontaneous spoken utterances and appears to work quite well. While the overall modular structure of the network is a symbolic design, the use of trainable, distributed networks allows for a certain level of generalization and fault tolerance. However, the flat parse lacks much of the information necessary to construct a full parse tree and does not confront some of the problems posed by ambiguities.

The Jain and Waibel (1990) model is essentially a localist, slot-based network, but it does incorporate learning and distributed representations at the word level. It consists of a series of layers representing words, phrases, clauses, and inter-clausal relationships. These layers are trained independently with specified targets and therefore involve only limited learned, distributed representations. The model is interesting in its ability to process inputs over time, producing expectations of sentence structure and dynamically revising hypotheses. However, it only has a fixed number of phrase and clause blocks and uses weight sharing to generalize learning across phrase blocks. This appears to cause a difficult tradeoff between proper generalization and over-generalization.

Another model to note is Stevenson's more recent parsing model (Stevenson, 1994; Stevenson & Merlo, 1997), which is based on $\overline{X}$ theory and is largely symbolic, but relies on activation-based competition mechanisms, as in localist network models, to resolve structural ambiguities. The XERIC model of Berg (1992) is also based on $\overline{X}$ theory but relies primarily on learned, distributed representations. XERIC combines a simple-recurrent network (SRN) (Elman, 1990) with a RAAM (Pollack, 1990) and is able to take words over time and produce a representation that can be decomposed into an $\overline{X}$ parse tree. This model has the advantage over localist methods that it can process unbounded sentences with only gradual degradation in performance. Although it was trained on a fairly simple grammar, the model is able to parse sentences with rather deep structure. Semantic information was not included in the original work, but it could theoretically be introduced into this model by using a microfeatural encoding for words at the input. Despite its successes, XERIC might not be considered an adequate cognitive model because its hierarchical training procedure, like that for the RAAM, requires considerable memory and symbolic control.

Henderson (1994a, 1994b, 1996) described a localist, non-learning connectionist parser based on *temporal synchrony variable binding* (TSVB) and inspired by symbolic parsing theories. The main idea behind TSVB is that variable bindings, such as the bindings of constituents to thematic roles, can be represented by synchronous firing of constituent and role representations. The use of temporal synchrony, rather than something like binding units, reduces the need for duplicate structure and permits greater generalization. Henderson argued that the overall architecture is biologically well-motivated. The model does not itself construct an entire parse tree, but a sequence of tree fragments with sufficient information to enable their recombination into a complete tree. Because it is a deterministic parser, never backtracking on its commitments, and because it is unable to represent disjunctions of interpretations, it is likely that this model would have great difficulty with ambiguous sentences and suffer from an overly strong garden-path effect. The main drawback of the model is that it is primarily a connectionist implementation of a symbolic algorithm and lacks many of the advantages of connectionist networks, including the ability to learn and make use of multiple weak constraints.

Henderson and Lane (1998) and Lane and Henderson

(1998) described an extension of the TSVB approach, known as a *simple synchrony network*, that can learn to parse sentences pre-encoded as parts-of-speech. The network takes the part of speech tags for the sentence constituents as input and is trained to produce the parse tree fragment of any constituent seen so far when that constituent is queried. The network was able to learn to parse a corpus of written English to a reasonable degree of proficiency. It is worth noting that TSVB may be identical in practice to the query mechanisms used in St. John and McClelland (1992) and Rohde (2002).

Finally, Harm, Thornton, and MacDonald (2000) were interested in how semantic and statistical regularities affect the parsing process. To begin to address this, they focused on the parsing of ambiguous N N or N V phrases such as “the desert trains.” Harm et al. trained a fully recurrent network (Pearlmutter, 1989) to process potentially ambiguous three-word phrases. As each word was presented, the network mapped from a distributed representation of the word’s form to a distributed representation of its meaning. Also present in the output was an indication of whether the phrase was an NP or an NP followed by a verb. Although this model is quite limited in its scope, it succeeded in demonstrating the desired sensitivity to a variety of factors, including structural constraints, pragmatic constraints, and lexical frequency and semantic biases.

Aside from the fact that they provide no account of language acquisition, symbolic, localist, and hybrid parsing models that cannot learn are not readily adapted to new or more complex languages and are generally insensitive to semantic constraints and other important sources of information in parsing. On the hand, parsing models that involve learned, distributed representations generally require teaching signals in the form of explicit parses, which are not thought to be available to language learners.

3.2 Word prediction

An alternative to parsing is the more basic, but nonetheless quite difficult, task of word prediction. Word prediction is a surprisingly useful ability. It can be the foundation for a *language model*, which predicts the likelihood that a particular utterance will occur in the language and which is a principal component of most speech recognition systems. The ability to predict accurately is sufficient to generate the language, and it thus indicates “weak” knowledge of the grammar underlying the language. Some of the most well-known and successful connectionist models of sentence processing are those that perform word prediction.

Elman (1990, 1991b, 1993) pioneered the use of simple-recurrent networks (SRNs), also called Elman networks for character and word prediction. Elman (1990) applied an SRN to letter prediction in a concatenated sequence of words, demonstrating that the network could potentially learn to detect word boundaries by identifying locations of high entropy, where the prediction is difficult. This work suggests that prediction might be a primary mechanism used by infants to learn word segmentation. Elman then extended the model to word prediction in a language of simple sentences. Representations of words that developed at the network’s hidden layer could be clustered to produce a reasonable classification of words syntactically and semantically. This indicates that much of the basic knowledge required for parsing and comprehension could be extracted from the child’s input by a prediction mechanism.

Elman (1991b) further extended the model to process sentences that potentially involve multiple embedded clauses. The main goal of this work was to demonstrate that networks are capable of learning to represent complex, hierarchical structure. As Elman put it, “The important result of the…work is to suggest that the sensitivity to context which is characteristic of many connectionist models, and which is built-in to the architecture of [SRNs], does not preclude the ability to capture generalizations which are at a high level of abstraction” (p. 220).

A second major outcome of the work was the finding that the networks were only able to learn corpora of mostly complex sentences if they first began training on simple sentences before gradually advancing to a higher proportion of complex ones. This was developed further in Elman (1993), where it was shown that the networks could also learn well if their memory spans were initially hindered and then gradually allowed to improve. This finding was thought to be particularly important as it accorded with Newport’s “less-is-more” hypothesis: that a child’s limited cognitive abilities may actually be a critical factor in enabling her to, ultimately, learn a first or second language to a greater degree of fluency than can an adult (Newport, 1990; Goldowsky & Newport, 1993).

Although these findings were influential and appeared to have important implications for human language learning, we re-examined them and discovered that manipulating the training environment or memory span of the networks does not always facilitate learning and can, in fact, be harmful (Rohde & Plaut, 1997, 1999). These studies used a similar network to Elman’s but a range of languages that differed in their statistical, but not syntactic, properties. The primary finding was that using initially simplified inputs was, in most cases, a significant hindrance to the networks. This was particularly true as the languages were made more natural through the introduction of semantic constraints. Memory impairments of the sort used by Elman, on the other hand, actually seem to have little effect on the learning of the network.

Our explanation for this was based on the fact that recurrent networks naturally begin with poor memory which they must gradually learn to use as they are exposed to the environment. The network therefore tends to learn simple relationships first because it does not yet have the representational capacity to handle more complex ones. Thus, Elman's staged memory impairments tend to have little effect because they simply mirror the natural development of memory. As argued in Rohde and Plaut (in press), "We believe that the cognitive limitations of children are only advantageous for language acquisition to the extent that they are symptomatic of a system that is unorganized and inexperienced but possesses great flexibility and potential for future adaptation, growth and specialization."

Weckerly and Elman (1992) used a similar SRN model and focused specifically on the difficulty of right-branching versus center-embedded sentences. They found that, in accordance with behavioral data, the SRN showed a preference for sentences involving double right-branching, subject-extracted relative clauses over those with double center-embedded, object-extracted clauses. Furthermore, the network was able to make use of semantic constraints to facilitate word prediction in center-embedded sentences.

Christiansen (1994) extended the language used by Elman (1991b) to include prepositional phrases, left recursive genitives, conjunction of noun phrases, and sentential complements. One version of the grammar could produce center-embedded sentences and a second version cross-dependencies. In general, the networks performed rather well on these languages and exhibited behaviors that largely reflect human comprehension performance on similar sentences. Christiansen and Chater (1999) further extended these results and provided more detailed comparisons with human performance.

Finally, Tabor, Juliano, and Tanenhaus (1997) (see also Tabor and Tanenhaus (1999)) performed a number of experiments comparing human and network reading times on sentences involving structural ambiguities. Reading times were elicited from an SRN using a novel "dynamical system" analysis. Essentially, the hidden representations that appear in the network at various stages in processing sentences are plotted in a high-dimensional space. These points are treated as masses that exhibit a gravitational force. To determine the reading time of the network on a particular word, the network's hidden representation for that word is plotted in the high-dimensional space and then allowed to gravitate among the attractors until a stable state is reached, with the settling time taken as an indicator of reading time. Although this test-mass settling process was intended to be a proxy for a true dynamical system that actually settles into a stable state, no experiments were performed to demonstrate that this is a reasonable simplification of such a model.

3.3 Comprehension

Comprehension models are those that go beyond parsing to producing a representation of the meaning of a sentence, given its surface form. There are, in fact, relatively few comprehension models in the literature. This may be due largely to the difficulty of representing and processing semantic information. Concept and phrase meanings involve subtle aspects that cannot easily be captured in a symbolic or localist system and do not interact in a cleanly combinatorial fashion. Furthermore, systems able to manipulate such information do not lend themselves to top-down design and are better constructed with learning methods. Therefore, comprehension has largely been the domain of distributed, connectionist models.

Hinton (1981) discussed one way in which semantic information and associations could be stored and recalled using distributed representations, and he pointed out some of the advantages this has over traditional localist semantic networks and over static distributed representations. A principal advantage is that associations formed between items may automatically generalize to semantically similar items. This work may have influenced, directly or indirectly, many subsequent connectionist models of semantics.

One such effort is the well-known model of McClelland and Kawamoto (1986). While it does not derive fully structured representations of sentence meaning, this model produces thematic case role assignments, which are thought to be an important step in comprehension. The proper assignment of case roles does not simply depend on word order but also involves considerations of word meaning, inflectional morphology, and context. The model uses stochastic units and a single layer of weights that is trained using the perceptron convergence rule and learns to map from the semantic features of the three or four main constituents of the sentence to the semantic representations for the fillers of up to four thematic roles: agent, patient, instrument, and modifier. The model is able to resolve lexical and structural ambiguities, handle shades of meaning, and generalize to novel words. However, as the authors acknowledged, this was just a first step which greatly simplified the problem of sentence comprehension. The use of static input representations does not allow the network to process words over time and results in a hard limit on the complexity of sentences that can be handled. In particular, this model would be unable to represent multi-clause sentences without substantial changes.

Perhaps the best known model of sentence comprehension is the later work of McClelland, St. John, and Taraban (1989) and St. John and McClelland (1992). These papers described a model that shares many of the goals of the McClelland and Kawamoto (1986) work but extends the framework to produce a changing interpretation as each constituent is received and to allow the learning of distributed hidden representations of phrase and sentence meaning. The input half of the model is an SRN that learns to use a sequence of phrase components to compile a single message representation, known as the sentence gestalt, in the form of a trainable hidden layer. The output half of the model was trained to answer questions about the sentence in the form of a probe. When probed with a constituent, the network must respond with the thematic role played by that constituent, or when probed with a role, the network produces the constituent that fills that role. During training, the error that derives from the answers to these probes is backpropagated through the network to influence the formation of the sentence gestalt.

The St. John and McClelland model successfully exhibited many interesting behaviors including the ability to make use of both syntactic and semantic clues to sentence meaning, revise its interpretations online, infer missing or vague constituents, use variable argument structure frames in both active and passive forms, and generalize its abilities to novel sentences. However, a major limitation of the model is that it, too, is not able to process multi-clause sentences, which are of considerable interest in the study of language. Nevertheless, the St. John and McClelland model remains a key inspiration for the work discussed in this paper.

One hindrance to the development of sentence comprehension models has been the difficulty of specifying adequate meaning representations of concepts and sentences. One solution adopted by Allen (1988), St. John (1992) and Noelle and Cottrell (1995) is to avoid specifying meanings by focusing on language learning in the service of a task. By grounding language in this way, the model can be trained to respond to linguistic inputs by performing an appropriate action. For example, St. John (1992) trained a simple-recurrent network to take a description of a scene and a sentence describing a particular object in the scene and identify the object to which the sentence refers. The model is able to handle fairly complex inputs including relative clauses and prepositional phrases and can even handle human-produced sentences moderately well, but is otherwise severely limited in its scope.

Miikkulainen and Dyer (1989) trained a backpropagation network on the same sentences used in the McClelland and Kawamoto (1986) study. The network learned to map from a static representation of the words in the sentence to a representation of the case role assignments. The principal difference between this and the earlier study is that McClelland and Kawamoto hand-designed feature-based distributed representations for words while the Miikkulainen and Dyer network learned the word representations using a representation updating, storage and retrieval mechanism. The method was later extended to a simple-recurrent network which accepts the same sentences encoded sequentially (Miikkulainen & Dyer, 1990).

Miikkulainen (1990) applied a modular architecture to comprehending and producing sentences with relative clauses. The sentences were composed of noun-verb or noun-verb-noun clauses, separated by commas. The first module maps from a sequence of words drawn from a single clause, or part of a clause if it contains an embedding, to a slot-based representation of the meaning. A second network maps from a sequence of clause frames to a static representation of all the frames in the sentence. Two other networks perform the inverse mappings. The system was able to successfully encode and reproduce sentences constructed from a very limited vocabulary. The use of a slot-filler representation for sentence meaning places a hard constraint on the complexity of sentences that could be represented by this system. Another limitation is that it relies on markers to distinguish clause boundaries, thus preventing it from handling reduced-relative constructions, which lack relative pronouns. Nevertheless this appears to have been the first connectionist comprehension model able to process complex sentences.

3.4 Production

Sentence production is a mapping from an intended meaning to a sequence of words or sounds. Production involves such issues as choosing words to convey the appropriate message, selecting the correct morphemes to obey syntactic and agreement constraints, and modeling the listener's knowledge to allow the speaker to avoid redundancy, provide an appropriate level of information, and produce syntactic forms and prosodic cues that emphasize important parts of the utterance and avoid ambiguity. Sentence production has received much less attention than parsing in the symbolist community. Producing the appropriate phrasing depends on sensitivity to nuances of meaning that are difficult to capture in a symbolic system (Ward, 1991). Thus, some researchers have begun turning to connectionist approaches to modeling production. However, most connectionist language production models have so far been restricted to the word level, dealing with lexical access and phoneme production, rather than sentence-level phenomena (Dell, 1986; O'Seaghdha, Dell, Peterson, & Juliano, 1992; Harley, 1993; Dell, Juliano, & Govindjee, 1993). This section considers the most notable sentence production networks.

There have been at least three major localist sentence production networks. Kalita and Shastri (1987, 1994) focused on the problem of producing the words in a sentence given the thematic role fillers and indications of the desired voice and tense. Their model, which is a rather complex localist network, is able to produce simple SVO sentences in active or passive voice and in several tenses.

In order to ensure that constituents are produced in the proper order, the model uses *sequencer* units to inhibit nodes once they have performed their duty. It is unlikely that this model could easily be extended to more complex sentences, particularly those with recursively nested structures.

Gasser (1988) (see also Gasser & Dyer, 1988) described a significantly more ambitious localist model that produces sentences using elaborate event schemas. The model, known as the Connectionist Lexical Memory, is based on interactive-activation principles. Bindings to syntactic roles are encoded with synchronized firing, as in TSVB (Henderson, 1994a). Sequencing is accomplished using start and end nodes for each phrase structure, which are somewhat similar to the sequencer units in Kalita and Shastri's model. Gasser's model is designed to account for a wide range of phenomena, including priming effects, speech errors, robustness given incomplete input or linguistic knowledge, flexibility in sequencing, and transfer of knowledge to a second language. However, the model was only applied to simple clauses and noun phrases and does not produce recursive structures involving long-distance dependencies. Again, it is not clear whether such a localist model could be scaled up to handle more complex sentences.

The third localist production model, by Ward (1991), was intended to be "more connectionist" than the previous attempts, relying on a truly interactive settling process and avoiding the need for binder units. One major limitation of the model, which may apply to the others as well, is that the network structures used to represent the intended meaning of the utterance are built on a sentence-by-sentence basis. Although the model is apparently able to produce a broader range of sentences, it is still unable to handle agreement, anaphor, and relative clauses. Ward acknowledged that a primary drawback of the model is the difficulty of extending it in all but the most trivial ways, and he recognized the need for a learning mechanism.

The localist models of sentence production, like their parsing cousins, suffer from an inability to learn or to handle complex structure without relying on redundancy or replication mechanisms. However, work on distributed connectionist models of production has been rather limited. We have already discussed the comprehension and production model of Miikkulainen (1990), which was trained to produce multi-clause sentences based on a slot-filler representation of its clauses. So far this work has been restricted to fairly simple domains. The nature of the representations used appears to limit the ability of the system to be scaled up to more natural languages. In earlier work, Kukich (1987) was interested in the ability of a network to learn to produce stock market reports given the day's activity. He trained one network to associate units of meaning, or sememes, with morphemes and another network to re-order morphemes. The output of the first network was an unordered set of word stems and suffixes, which could be produced accurately 75% of the time. The morpheme-ordering network did not actually produce morphemes sequentially but used a slot-based encoding of order. The results of these simulations left considerable room for improvement but were encouraging given the early state of connectionism.

More recently, Dell, Chang, and Griffin (1999) were specifically interested in the phenomenon of structural priming, which leads speakers to preferentially produce sentences of a particular form, such as passive rather than active voice, if they have recently heard or produced sentences of similar form. Dell et al. hypothesized that the mechanism that results in structural priming is the same procedure used to learn production. Their model takes a representation of the sentence's propositional content and produces the words in the sentence sequentially. Propositional content is encoded using a slot-based representation of the clause constituents, and the model was thus able to produce only simple sentences.

The model was trained to produce either active or passive sentences based on whether the agent or patient received greater emphasis. It was also able to convey recipients using a prepositional phrase or a dative. The model learned to produce sentences with 94% of the words, or about 75% of sentences, being correct, and thus was not as accurate as one might hope. However, the model was able to match human structural priming data quite well. The main limitations of this model are that it was applied only to simple sentences and did not learn distributed context representations.

3.5 The CSCP model

Other than prediction networks, which avoid the issue of meaning entirely, no connectionist sentence processing models discussed thus far have exhibited all of the main properties necessary to provide a plausible account of natural language acquisition. These include the ability to learn a grammar, to process a sentence sequentially, to represent complex, multi-clause sentences, and to be naturally extendable to languages outside of the domain originally addressed by the designer. The Connectionist Sentence Comprehension and Production (CSCP) model of Rohde (2002) was developed to address these limitations.

The CSCP model is essentially a large-scale simple-recurrent network able to perform both comprehension and production of complex, multi-clausal sentences using a subset of English constrained to match English in terms of its distributional properties. This language involves such complexities as multiple verb tenses and voices, adverbs and adjectives, prepositional phrase, relative and subordinate clauses, and sentential complements. Sentence meanings are composed of a set of propositions encoded using distributed, featural representations. One portion of the network, the semantic system, learns to compress a sequence of these propositions into a single, static representation of the meaning of the sentence under the pressure to answer fill-in-the-blank questions about the stored propositions as in the St. John and McClelland (1992) model.

The actual comprehension portion of the system then learns to receive a sequence of words, encoded in a distributed phonological representation, and output a representation of the sentence meaning that can be decoded by the semantic system. Although the model is provided with knowledge about sentence meanings and word segmentation, it must learn to induce the syntax of the language. Building on the work of Elman (1991b), the comprehension system is simultaneously trained to predict the next word in the sentence. At times, the meaning of the sentence is provided to the model in advance, enabling it to generate much more accurate predictions. This, ultimately, serves as the basis for the model's sentence production. In order to produce a sentence, the message layer is clamped to the correct meaning and the model predicts the first word in the sentence. The most strongly predicted word is selected and fed back into the model's comprehension input and it proceeds to produce the next word, and so on.

Therefore, the model's comprehension and production mechanisms are tightly integrated and rely on many of the same processes. An intrinsic claim is that language production is learned primarily through formulating implicit predictions while listening and attempting to comprehend sentences in one's language. The CSCP model has been extensively tested on a variety of tasks, including the processing of lexical and structural ambiguities, and a range of unambiguous sentence types. It is able to replicate many key aspects of human sentence processing, including sensitivity to structural frequency, verb argument structure preferences, inflectional morphology, locality effects, and semantic plausibility. In production, it also demonstrates structural priming and number agreement attraction. The model's sensitivity to particular statistical factors, and the representations on which this depends, arise naturally from the constraints of its connectionist architecture as it learns to perform the tasks of comprehension and production as best it can.

Acknowledgements

This research was supported by NIMH Program Project Grant MH47566 (J. McClelland, PI), and by NIMH NRSA 1-F32-MH65105-01. Correspondence regarding this article may be sent either to Douglas Rohde (dr@tedlab.mit.edu), MIT, NE20-437E, 3 Cambridge Center, Cambridge, MA 02139, USA, or to David Plaut (plaut@cmu.edu), Mellon Institute 115–CNBC, 4400 Fifth Avenue, Pittsburgh, PA 15213–2683, USA.

References

Allen, R. B. (1988). Sequential connectionist networks for answering simple questions about a microworld. In *Proceedings of the 10th annual conference of the Cognitive Science Society* (pp. 489–495). Hillsdale, NJ: Lawrence Erlbaum Associates.

Berg, G. (1992). A connectionist parser with recursive sentence structure and lexical disambiguation. In *Proceedings of the 10th National Conference on Artificial Intelligence* (pp. 32–37). San Jose, CA: AAAI.

Charniak, E., & Santos, E. (1987). A connectionist context-free parser which is not context-free, but then it is not really connectionist either. In *Proceedings of the 9th annual conference of the Cognitive Science Society* (pp. 70–77). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chomsky, N. (1957). *Syntactic structure.* The Hague, The Netherlands: Mouton.

Christiansen, M. H. (1994). *Infinite languages, finite minds: Connectionism, learning, and linguistic structure.* Unpublished doctoral dissertation, University of Edinburgh.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204-256.

Cottrell, G. W. (1985a). Connectionist parsing. In *Proceedings of the 7th annual conference of the Cognitive Science Society* (pp. 201–211). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cottrell, G. W. (1985b). *A connectionist approach to word sense disambiguation.* Unpublished doctoral dissertation, Department of Computer Science, University of Rochester, Rochester, NY.

Cottrell, G. W., & Plunkett, K. (1995). Acquiring the mapping from meaning to sounds. *Connection Science Journal of Neural Computing, Artificial Intelligence and Cognitive Research*, *6*(4), 379–412.

Dell, G. S. (1986). A spreading activation theory of retrieval in language production. *Psychological Review*, *93*, 283–321.

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. In M. H. Christiansen, N. Chater, & M. S. Seidenberg (Eds.), *Special Issue of Cognitive Science: Connectionist Models of Human Language Processing: Progress and Prospects* (Vol. 23). Cognitive Science.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and context in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, *17*, 149–195.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*, 801–838.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J. L. (1991a). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195-225.

Elman, J. L. (1991b). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Fahlman, S. E., Hinton, G. E., & Sejnowski, T. J. (1983). Massively parallel architectures for AI: NETL, Thistle, and Boltzmann machines. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 109–113). Washington.

Fanty, M. (1985). *Context-free parsing in connectionist networks* (Tech. Rep. No. TR-174). Rochester, NY: University of Rochester, Computer Science Department.

Fodor, J. A. (1983). *Modularity of mind.* Cambridge, MA: MIT Press.

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies.* Bloomington, IN: Indiana University Linguistics Club.

Frost, R., Deutsch, A., & Forster, K. I. (2000). Decomposing morphologically complex words in a nonlinear morphology. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 751-765.

Gasser, M., & Dyer, M. G. (1988). Sequencing in a connectionist model of language processing. In *COLING Budapest: Proceedings of the 12th International Conference on Computational Linguistics* (pp. 185–190). Budapest.

Gasser, M. E. (1988). *A connectionist model of sentence generation in a first and second language.* Unpublished doctoral dissertation, Computer Science Department, University of California, Los Angeles, CA. (TP UCLA-AI-88-13)

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*, 447-474.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: the less is more hypothesis. In E. Clark (Ed.), *The proceedings of the 24th annual Child Language Research Forum* (pp. 124–138). Stanford, CA: Center for the Study of Language and Information.

Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (submitted). *Graded semantic and phonological effects in priming: Evidence for a distributed connectionist approach to morphology.* (Manuscript submitted for publication, July 1999.)

Hare, M., & Elman, J. (1995). Learning and morphological change. *Cognition*, *56*, 61–98.

Harley, T. A. (1993). Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes*, *8*, 291–309.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*(3), 491-528.

Harm, M. W., & Seidenberg, M. S. (2001). Are there orthographic impairments in phonological dyslexia? *Cognitive Neuropsychology*, *18*, 71-92.

Harm, M. W., Thornton, R., & MacDonald, M. C. (2000). A distributed, large scale connectionist model of the interaction of lexical and semantic constraints in syntactic ambiguity resolution [Abstract]. In *Proceedings of the 13th annual CUNY Conference on Human Sentence Processing.* La Jolla, CA.

Henderson, J. B. (1994a). Connectionist syntactic parsing using temporal variable binding. *Journal of Psycholinguistic Research*, *23*(5), 353–379.

Henderson, J. B. (1994b). *Description based parsing in a connectionist network.* Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, PA.

Henderson, J. B. (1996). A connectionist architecture with inherent systematicity. In *Proceedings of the 18th annual conference of the Cognitive Science Society* (pp. 574–579). Hillsdale, NJ: Lawrence Erlbaum Associates.

Henderson, J. B., & Lane, P. C. R. (1998). A connectionist architecture for learning to parse. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th annual meeting of the Association for Computational Linguistics (COLING-ACL'98).* University of Montreal, Canada.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 161–187). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoeffner, J. H., & McClelland, J. L. (1993). Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation. In E. V. Clark (Ed.), *Proceedings of the 25th Annual Child Language Research Forum* (p. 38-49). Stanford, CA: Center for the Study of Language and Information.

Howells, T. (1988). VITAL: A connectionist parser. In *Proceedings of the 10th annual conference of the Cognitive Science Society* (pp. 18–25). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jain, A. N., & Waibel, A. H. (1990). Incremental parsing by modular recurrent connectionist networks. In D. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 364–371). San Mateo, CA: Morgan Kaufmann.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Science, USA*, *96*, 7592-7597.

Jordan, M. I., & Rumelhart, R. A. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*, 307–354.

Kalita, J., & Shastri, L. (1987). Generation of simple sentences in English using the connectionist model of computation. In *Proceedings of the 9th annual conference of the Cognitive Science Society* (pp. 555–565). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kalita, J., & Shastri, L. (1994). A connectionist approach to generation of simple sentences and word choice. In G. Adriaens & U. Hahn (Eds.), *Parallel natural language processing* (pp. 395–420). Norwood, NJ: Ablex Publishing.

Kukich, K. (1987). Where do phrases come from: Some preliminary experiments in connectionist phrase generation. In G. Kempen (Ed.), *Natural language generation: New results in artificial intelligence, psychology and linguistics* (pp. 405–421). Boston, Dordrecht: Martinus Nijhoff Publishers.

Kwasny, S. C., & Faisal, K. A. (1990). Connectionism and determinism in a syntactic parser. *Connection Science*, *2*, 63–82.

Lane, P. C. R., & Henderson, J. B. (1998). Simple synchrony networks: Learning to parse natural language with temporal synchrony variable binding. In *Icann.* Skövde, Sweden.

Locke, J. L. (1983). *Phonological acquisition and change.* New York: Academic Press.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, *40*, 121-153.

Marchman, V. A. (1993). Constraints on plasticity in a connectionist model of the English past tense. *Journal of Cognitive Neuroscience*, *5*, 215–234.

Marcus, M. P. (1980). *A theory of syntactic recognition for natural language.* Cambridge, MA: MIT Press.

McClelland, J. L., & Kawamoto, A. H. (1986). Mechanisms of sentence processing: Assigning roles to constituents of sentences. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 272–325). Cambridge, MA: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*(5), 375-407.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group the (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models.* Cambridge, MA: MIT Press.

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

McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes.* Oxford, UK: Oxford University Press.

Menn, L., & Stoel-Gammon, C. (1995). Phonological development. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (p. 335-359). Oxford: Blackwell.

Miikkulainen, R. (1990). A PDP architecture for processing sentences with relative clauses. In *COLING-90: Papers presented to the 13th International Conference on Computational Linguistics* (pp. 3/201–206). Helsinki.

Miikkulainen, R., & Dyer, M. (1990). *Natural language processing with modular neural networks and distributed lexicon* (Tech. Rep. No. CSD-900001). Los Angeles, CA: Computer Science Department, University of California.

Miikkulainen, R., & Dyer, M. G. (1989). Encoding input/output representations in connectionist cognitive systems. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. 347–356). Los Altos, CA: Morgan Kaufman.

Nakagawa, H., & Mori, T. (1988). A parser based on connectionist model. In *COLING Budapest: Proceedings of the 12th International Conference on Computational Linguistics* (pp. 454–458). Budapest.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *34*, 11–28.

Noelle, D. C., & Cottrell, G. W. (1995). A connectionist model of instruction following. In *Proceedings of the 17th annual conference of the Cognitive Science Society* (pp. 369–374). Hillsdale, NJ: Lawrence Erlbaum Associates.

O'Seaghdha, P. G., Dell, G. S., Peterson, R. R., & Juliano, C. (1992). Models of form-related priming in comprehension and production. In R. G. Reilly & N. E. Sharkey (Eds.), *Connectionist approaches to natural language processing* (pp. 373–408). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, *1*, 263–269.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., & Jordan, M. I. (1995). Goal-based speech motor control: A theoretical framework and some preliminary data. *Journal of Phonetics*, *23*, 23-35.

Pinker, S. (1999). *Words and rules: The ingredients of language.* New York: Basic Books.

Pinker, S., & Mehler, J. (Eds.). (1988). *Connections and symbols.* Cambridge, MA: MIT Press.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*(4/5), 445-485.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381–415). Mahwah, NJ: Lawrence Erlbaum Associates.

Plaut, D. C., Seidenberg, M. S., McClelland, J. L., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Plunkett, K., & Marchman, V. A. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, *38*, 43-102.

Plunkett, K., & Marchman, V. A. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, *48*(1), 21-69.

Plunkett, K., & Marchman, V. A. (1996). Learning from a connectionist model of the acquisition of the English past tense. *Cognition*, *61*(3), 299-308.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*, 77–105.

Quinlan, P. (1991). *Connectionism and psychology: A psychological perspective on new connectionist research.* Chicago: University of Chicago Press.

Rager, J. E. (1992). Self-correcting connectionist parsing. In R. G. Reilly & N. E. Sharkey (Eds.), *Connectionist approaches to natural language processing* (pp. 143–167). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rohde, D. L. T. (2002). *A connectionist model of sentence comprehension and production.* Unpublished doctoral dissertation, Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA.

Rohde, D. L. T., & Plaut, D. C. (1997). Simple recurrent networks and natural language: How important is starting small? In *Proceedings of the 19th annual conference of the Cognitive Science Society* (pp. 656–661). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*(1), 67–109.

Rohde, D. L. T., & Plaut, D. C. (in press). Less is less in language acquisition. In P. Quinlan (Ed.), *Studies in developmental psychology: Connectionist models of development.* Charles Hume.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (p. 216-271). Cambridge, MA: MIT Press.

Seidenberg, M. S. (1997). Language acquistion and use: Learning and applying probabilistic constraints. *Science*, *275*(5306), 1599-1603.

Seidenberg, M. S., & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, *4*, 353-361.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523-568.

Selman, B., & Hirst, G. (1985). Connectionist parsing. In *Proceedings of the 7th annual conference of the Cognitive Science Society* (pp. 212–221). Hillsdale, NJ: Lawrence Erlbaum Associates.

Selman, B., & Hirst, G. (1994). Parsing as an energy minimization problem. In G. Adriaens & U. Hahn (Eds.), *Parallel natural language processing* (pp. 238–254). Norwood, NJ: Ablex Publishing.

Small, S., Cottrell, G., & Shastri, L. (1982). Toward connectionist parsing. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 247–250). Pittsburgh, PA: AAAI.

Stevenson, S. (1994). *A competitive attachment model for resolving syntactic ambiguities in natural language parsing.* Unpublished doctoral dissertation, Department of Computer Science, University of Maryland.

Stevenson, S., & Merlo, P. (1997). Lexical structure and parsing complexity. *Language and Cognitive Processes*, *12*, 349–399.

St. John, M. F. (1992). Learning language in the service of a task. In *Proceedings of the 14th annual conference of the Cognitive Science Society* (pp. 271–276). Hillsdale, NJ: Lawrence Erlbaum Associates.

St. John, M. F., & McClelland, J. L. (1992). Parallel constraint satisfaction as a comprehension mechanism. In R. G. Reilly & N. E. Sharkey (Eds.), *Connectionist approaches to natural language processing* (pp. 97–136). Hillsdale, NJ: Lawrence Erlbaum Associates.

Studdert-Kennedy, M. (1993). Discovering phonetic function. *Journal of Phonetics*, *21*, 147-155.

Tabor, W., Juliano, C., & Tanenhaus, M. K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, *12*(2/3), 211–271.

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing. In M. H. Christiansen, N. Chater, & M. S. Seidenberg (Eds.), *Special Issue of Cognitive Science: Connectionist Models of Human Language Processing: Progress and Prospects* (Vol. 23). Cognitive Science.

Tepper, J. A., Powell, H. M., & Palmer-Brown, D. (2001). Corpus-based connectionist parsing. In *The Second Workshop on Natural Language Processing and Neural Networks (NLPNN2001).* Tokyo.

Vihman, M. M. (1996). *Phonological development: The origins of language in the child.* Oxford: Blackwell.

Waltz, D. L., & Pollack, J. B. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, *9*, 51–74.

Ward, N. (1991). *A flexible, parallel model of natural language generation.* Unpublished doctoral dissertation, Computer Science Division, University of California, Berkeley, CA. (UCB/CSD 91/629)

Weber, V., & Wermter, S. (1996). Using hybrid connectionist learning for speech/language analysis. In S. Wermter, E. Riloff, & G. Scheler (Eds.), *Lecture notes in artificial intelligence 1040: Connectionist, statistical, and symbolic approaches to learning for natural language processing* (pp. 87–101). Berlin: Springer-Verlag.

Weckerly, J., & Elman, J. L. (1992). A PDP approach to processing center-embedded sentences. In *Proceedings of the 14th annual conference of the Cognitive Science Society* (pp. 414–419). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wermter, S., & Weber, V. (1994). Learning fault-tolerant speech parsing with screen. In *Proceedings of the 12th National Conference on Artificial Intelligence* (pp. 670–675). Seattle, WA: AAAI.

Wermter, S., & Weber, V. (1997). SCREEN: Learning a flat syntactic and semantic spoken language analysis using artificial neural networks. *Journal of Artificial Intelligence Research*, *6*, 35–85.

Younger, D. H. (1967). Recognition and parsing of context-free languages in time n^3. *Information and Control*, *10*(2), 189–208.