Connectionist Sentence Processing in Perspective*

Mark Steedman

Draft 2.1, May 5, 1999

*To appear in *Cognitive Science*. Thanks to Gerry Altmann, Nick Chater, Morten Christiansen, Michael Collins, Geoff Hinton, Mark Liberman, Jane Neumann, and the referees for comments and advice. The research was supported in part by NSF grant nos. IRI91-17110, IRI95-04372, ARPA grant no. N66001-94-C6043, ARO grant no. AAH04-94-G0426, and ESRC Award Number M/423/28/4002.

Connectionist Sentence Processing in Perspective

Mark Steedman, University of Pennsylvania and University of Edinburgh

Abstract

The emphasis in the connectionist sentence-processing literature on distributed representation and emergence of grammar from such systems can easily obscure the often close relations between connectionist and symbolist systems. This paper argues that the Simple Recurrent Network (SRN) models proposed by Jordan (1990) and Elman (1990) are more directly related to stochastic Part-of-Speech (POS) Taggers than to parsers or grammars as such, while auto-associative memory models of the kind pioneered by Longuet-Higgins, Willshaw, Pollack and others may be useful for grammar induction from a network-based conceptual structure as well as for structure-building.

These observations suggest some interesting new directions for specifically connectionist sentence processing research, including more efficient representations for finite state machines, and acquisition devices based on a distinctively connectionist basis for grounded symbolist conceptual structure.

1 Introduction

As many papers in this special issue attest, an active and constructive dialogue about processing at the level of spoken and written words, and about the acquisition of related systems such as phonology, morphology and the lexicon, is going on between symbolic or rule-oriented theorists and connectionist or neurally-oriented theorists (see Christiansen and Chater, this issue, for an overview) . There seems to be much less dialogue between symbolic and rule-based approaches to syntactic analysis, despite genuine efforts to reconcile these positions on the connectionist side, for example by Hinton (1990c), Smolensky (1990) and this issue, and others collected in Hinton (1990a). This paper is a reciprocal attempt from the symbolist side to get such a dialogue going on the basis of the contributions collected here.

The traditional rule-based view of syntactic processing divides the problem among various modules. One fairly generally applicable way of doing this is to distinguish a *grammar*, characterized by syntactic and semantic rules of certain classes and a related characteristic automaton, an *algorithm* characterized by properties like the order in which rules are applied to the string, whether bottom-up or top-down, and by certain memory resources, such as those used in building structure and the charts or tables used in parsers based on dynamic programming, and an *oracle* or decision criterion for deciding which rule to apply the algorithm to in cases where there is more than one possibility.

In any given theoretical presentation or implementation, these modules may be combined, but in rule-based theories they can usually be distinguished in functional terms. The fact that they are in that sense distinct modules does not of course imply that the corresponding computations must be carried out in a series of chronologically distinct phases: for example it is quite possible to construct systems in which the oracle can call on the results of semantic interpretation in mid-parse, while grammatical analysis and the algorithm are still under way.

Connectionism is no more intrinsically non-modular than any other approach, and many connectionists including some represented here have explicitly endorsed modular architectures of various kinds. Nevertheless, the emphasis in the connectionist literature on distributed representation and "emergence" of rule-like behavior from such systems has sometimes made it hard for connectionists and symbolists alike to recognize the often close relations between their respective systems.

Part of the difficulty in reconciling the two stems from the involvement of two rather different views of the role of grammar in processing. The "performance" grammar used by the three-module processor described above can in theory be quite different from the grammar that linguists identify as the "competence" grammar. The linguists' grammar is usually one whose derivation structures are closely related to their intuitions about the interpretation of sentences. (Since we have no direct access to interpretations, the practical criteria for choosing one linguistic analysis over another are usually described in rather different terms by linguists like Chomsky, but in fact this is what it comes down to.)

Interpretations, and hence the linguistic competence grammars that (how ever imperfectly) reflect them, have a number of important properties. Most importantly, interpretations are *compositional*, which means that they are recursively defined solely in terms of the interpretations of their parts. This means that to know the meaning/grammatical category of a predicate like *walks* is to know the meaning/grammaticality of the proposition that results from applying that predicate to any argument of the appropriate type. Similarly, to know the meaning/grammatical category of a verb like *deny* is to know the meaning/grammaticality of denying any proposition in the language including those involving the verb *deny*. Compositionality entails a property that Fodor &

Pylyshyn (1988) proposed as a test for grammar-induction or emergence which they called *systematicity*, which means that a system cannot be claimed to embody a grammar unless it can recognise the grammaticality and ultimately the interpretation of any sentence of the language, whether or not it has been encountered before—see Hadley (1994a,b) on definitions of this property of increasing levels of strictness up to "semantic" systematicity.

If a grammar accepts all and only the strings that some other grammar accepts, then the two are said to be weakly equivalent. Some early parsing programs used algorithms requiring grammars in a normal form that was not particularly congenial to linguists. They therefore used a weakly equivalent normal-form "covering grammar" to build structures that could be transduced into the structural representation required by the linguists. Modern parsers often compile large grammars into finite-state covering grammars (whose coverage must of course necessarily be incomplete) for reasons of efficiency—see Black (1989).

For reasons of evolutionary simplicity that will be elaborated below, it would actually be rather surprising to find that human processors used a covering grammar. Nevertheless it is possible in theory, and there are some properties of human processors that might appear to suggest that they do. In particular, there are well-known (if poorly understood) limitations on human abilities to process sentences involving center embedding. Since center-embedding is one of the properties of natural grammars that led Chomsky (1957) to claim that context-free grammars represented a lower bound on expressive power, it has sometimes been claimed that these limitations are evidence that human processors work with an incomplete finite-state covering grammar. (Of course, other explanations are possible. It might be the algorithm that is incomplete, perhaps because of memory limitations, or even that there is something about these constructions that irrevocably misleads the oracle.)

However, if a covering grammar is involved, it must be one that is capable of specifying the interpretation. That is not the same as saying that the derivations themselves must correspond to traditional syntactic structure. Linguists tend to think of syntactic derivation as surface structure-building, but it is equally possible to think of such structures (as computational linguists tend to) as implicit in the flow of control in a parser that incrementally builds the interpretation directly (the early Augmented Transition Network (ATN) parser of Woods 1970 had this character). Such derivations can be structurally quite unlike traditional grammarians' analyses.

Linguists (especially computational linguists) also usually think of interpretations as structures or logical forms, but as a matter of fact this is not strictly necessary either. As in the work of Montague (1970, 1974), it is possible in principle to regard logical form as no more than the flow of control in computing models, in the sense of that term used in model-theoretic semantics. However, it is important not to get carried away by this possibility. Model theory is really only good for proving very general properties of formal systems, such as soundness and completeness. In AI and natural language understanding we are usually interested in the *form* of a constructive proof that we can get to New York City, rather than the mere truth of that proposition, because such a proof tells us *how* to actually get there. This means that practical knowledge representation systems are almost always proof-theoretic, involving manipulation of formulae, rather than model-theoretic. The significance of this point is that, when faced with a connectionist processor, one must not only ask what grammar is implicit in it, but how it can be made to deliver logical forms of a kind that can be manipulated by the equivalent of rules like *Modus Ponens*.

A second source of difficulty in comparing connectionist and rule-based systems lies in the tendency of certain kinds of connectionist architecture to combine the roles of grammar and oracle or ambiguity-resolution device in a single representation distributed over a single set of hidden units. This can make it hard to know whether one should compare the systems as a whole, or regard the connectionist system as merely a disambiguating device—for example, as the analog of a Markovian part-of-speech (POS) filter in a standard parsing architecture. I shall argue below that some devices that have presented as sentence processors should be thought of in this more restricted sense, and that (if the problems of reliability and scalability inherent in mechanisms based on back-propagation can be overcome), this view offers a way forward to a kind of system that embodies both rules and sub-symbolic representations in a principled way.

2 Recurrent Networks and Finite State Disambiguators

The recurrent networks proposed by Jordan (1989) and Elman (1990) use an auxiliary bank of "state" or "context" units to store information about the previous state of an otherwise standard three-level feed-forward network using back-propagation to adjust a hidden layer of units. The recurrence consists of copying either the output units or the hidden units to the context units. The context units then provide some of the inputs to the hidden units. Jordan applies such a device to the coarticulation problem in speech, and uses context units to represent

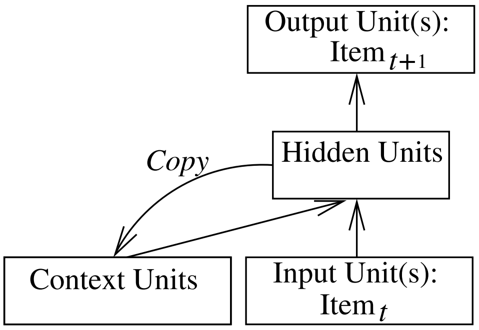

Figure 1
Architecture of the Simple Recurrent Network (SRN)

the preceding items directly by copying the output units from the previous cycle. Copying the hidden units, as in Elman's Simple Recurrent Network (SRN) shown in Figure 1, can represent more abstract properties of the preceding sequence. Interestingly, the context units can come to carry "echoes" of earlier states of the computation, and can thereby be used to represent quite distant dependencies between string elements. Elman applies SRNs to the problem of supervised grammar induction, in a rather indirect sense in which the device (since it does not build structure or exhibit internal states) can only reveal the implicit grammar indirectly, by predicting the next word or word category. This of course is a test related to weak equivalence of the implicit grammar.

Elman's early work was criticized by Hadley (1994a) for non-systematicity in the selection of test materials with respect to the claimed implicit grammar, so that in some cases the implicit grammar was not even proved weakly equivalent to the one that was initially claimed. However Christiansen & Chater (1994), Niklasson & van Gelder (1994), and the work by Allen (1997) discussed by Seidenberg & MacDonald in this issue show that SRNs can achieve systematicity in acquiring implicit grammars weakly equivalent to phrase structure fragments. Elman's work is also extended by Cleeremans (1993) and by Christiansen & Chater (in press), among many others. In particular, Christiansen has shown that SRNs can cover (finitely bounded sub-languages of) small context-free grammars with center embedding, and (similarly bounded sub-languages of) small trans-context free grammars including crossing dependencies of the kind notoriously characteristic of Germanic verb raising constructions (Bresnan, Kaplan, Peters, & Zaenen 1982; Huybregts 1984, Shieber 1985). There is no doubt that SRNs and other recurrent nets can approximate covering finite-state machines of this kind. (See Cleeremans 1993; Cleeremans, Servan-Schreiber, & McClelland 1995; Casey 1996.) The only obvious limitation on the approximation is that it appears to rapidly become less reliable for a given number of hidden units as the distance of the dependency increases. For this reason the precise place of such "graded state" automata in the automata-theoretic hierarchy is not entirely clear, but in practice they seem to be limited to finite state machines.)

It is important to recall that the sole task that the SRN is required to do is to predict the next word or word category at each point in the string. We know from work on symbolic finite-state models such as Hidden Markov Models (HMM) and part-of-speech (POS) taggers (Jelinek, 1976; Church, 1988; Merialdo, 1994; Brill, 1992) that such approximations can achieve very high accuracy—better than 95% precision—without having any claim whatsoever to embody the grammar itself. (To put this number in perspective, one should however recall that prediction of category on the basis of simple unigram frequency alone yields around 91% accuracy.) One of the surprising things about the recurrent network literature is that there is very little link to statistical computational linguistics, despite some early identifications of an equivalence relation by Williams & Hinton (1990) and Bridle (1992). Nevertheless, the comparison with POS taggers seems to be a relevant one.

The resemblance of SRNs to POS taggers is even closer among versions like that presented by Tabor & Tanenhaus (this issue) in which a principal components analysis of the hidden units is used to reveal implicit grammatical categories. It has even been suggested that the trajectory that a sequence of such categories defines through the high-dimensional space defined by the hidden units and/or principal components can be thought of as defining a meaning (Elman 1995), a claim that would begin to address the issue of semantic systematicity.

Although we shall see below that hidden units can legitimately be viewed as encoding certain kinds of semantic information, this last claim seems too strong. A mere sequence of words, even a sequence that is successfully disambiguated as to syntactic category and even word sense (as is reasonable to expect from a stochastic tagger), is not a sentence meaning, as is evident from the fact that the following string remains structurally and semantically globally ambiguous even when the lexical categories are unambiguously identified:

(1) Put the block in the box on the table.

We will defer until later the discussion of whether the further information that is needed to resolve the ambiguity is probabilistic or inferential, and whether the special graded nature of SRNs can capture the probabilistic alternative. The relevant point is that a significant component of the grammar induction and parsing problems as they are usually understood remains to be dealt with once n-gram based POS taggers have done their work, and the same appears to be true for the SRN as it is used in these experiments.

This observation should not be taken to deny that SRNs are useful as a component of sentence processors. SRNs and related devices may be a very good way of building stochastic part-of-speech disambiguators as an input to parsing proper. This seems to be the role that the SRN plays in the modular architectures of St. John & McClelland (1990), Berg (1992), Sharkey & Sharkey (1992) and Mikkulainen (1993, 1995) discussed below. Interestingly, Srinivas (1997) and Srinivas & Joshi (1994) have shown that increasing the set of POS tags to include the sub-categorization or domain of government information implicit in the lexicalist grammars discussed in section 5 increases the effectiveness of such devices, and Kim, Srinivas, & Trueswell (1998) show that SRNs can induce a tagger for such extended category sets. We shall also see below that SRNs can in principle be used to look at string contexts that extend beyond the sentential boundary. But none of these undoubted virtues suggest that it is correct to regard grammar itself as in any sense an emergent property of these devices as presently used.

It is also worth noting that distributed representations potentially allow exponentially greater efficiency in representation of stochastic finite state machines over those induced by HMMs (Williams & Hinton, 1990) (although to the extent that SRNs seriously exploit the potential of hidden units for efficiently distributing the representation of the FSM they approximate, it becomes correspondingly harder to see how to associate structure building operations of any kind with them directly.) It is not likely that the SRN itself will achieve such efficiency because of the trade-off that it makes between full back-propagation through time and on-line computation—see Pearlmutter (1995) for relevant discussion. But if POS tagging is what SRN is actually doing, then we may not want to interpret hidden unit states, but may rather prefer to exploit the efficient way in which they and some of the other devices discussed by Williams and Hinton can exploit information theoretic redundancy in text quite independently of grammar. For example, the Latent Semantic Analysis (LSA) program of Landauer & Dumais (1997) when trained on large volumes of text and tested on similarity judgements between words and passages showed very high correlations with human similarity judgements. More recent work in the framework using similarity measures between student essays (which LSA treats as unordered bags of words) and instructional text for the relevant domain yielded correlations with grades assigned to the essays by human assessors comparable to the correlation across human graders themselves (Landauer, Laham, Rehder, & Schreiner, 1997). However, despite its name and its very interesting performance, this result cannot be equated with the delivery of a meaning, as anyone who does the thought experiment of running the theory in the opposite direction to generate the student essays will agree.

3 Psychological Relevance of SRN

A number of studies have investigated the fit of SRNs to human parsing performance. The studies in this issue by Dell, Chang, and Griffin, and Tabor & Tanenhaus are examples of this kind of work. Tabor & Tanenhaus advance the theoretical model via an elaboration of the copying mechanism which the SRN uses to approximate Back-propagation Through Time (BTT, Rumelhart, Hinton, & Williams, 1986; Pearlmutter 1995), and by a "gravitational" analysis using attractors obtained from a principal components analysis of the patterns of activation on the hidden units to interpret them as FSM states. The authors point out that these can be viewed as parse hypotheses which can be mapped onto more traditional symbolic models (although as we have seen, the close relation between POS tagging and SRNs makes it likely that such a translation will in general be non-trivial, and that it may be seriously incomplete).

Tabor and Tanenhaus also provide a detailed comparison with the experimental findings of McRae, Spivey-Knowlton, & Tanenhaus (1997) concerning the influence of "thematic fit" of verbs and arguments in on-line sentence comprehension for minimal pairs of sentences like (2a,b), in which the misleadingly better thematic fit of *cop* as agent than as patient in an indicative reading of *arrested* in comparison to *crook* causes a temporary increase in processing load later in sentence (a), when that reading leads to no grammatical analysis.

(2) a. The cop arrested by the detective was guilty.
    b. The crook arrested by the detective was guilty.

The fit of the model in terms of predicting word by word processing effort as revealed by increased reading times is impressive. The authors are justified in their claim that the SRN can cover the same phenomenon as the structure-invoking "garden path" model (whose most recent incarnation is Frazier & Clifton 1996), without building any structure at all.

However, the claim that this amounts to the "emergence" of grammar from the SRN model seems premature. It remains the case that the only thing the SRN is doing is predicting the next category in the sentence. It is actually not in the least surprising that the SRN can do at least as well as the structural model. It is learning a finite state machine, and we have seen that FSMs can do very well at category prediction for homogeneous corpora. Earlier work by Tanenhaus' own group has shown that structural properties of sentences are confounded with frequency effects, and that when frequency is properly controlled, there is no evidence for any residual purely structural preference. (See Trueswell & Tanenhaus 1992; Trueswell, Tanenhaus, & Kello, 1993; Trueswell, Tanenhaus, & Garnsey, 1994; Spivey-Knowlton, Trueswell, and Tanenhaus, Spivey-Knowlton et al. 1993).

Nor does the success of the SRN in predicting processing difficulty on the basis of frequency information alone justify any claim that the human processor works on the same basis. Frequency is even more strongly confounded with semantic and pragmatic plausibility, since word transition probabilities are compiled from coherent text corpora. As Tanenhaus and colleague's earlier papers are careful to point out, their results therefore do not distinguish between a model of the processor like SRN that resolves ambiguity entirely probabilistically and one where ambiguity is resolved on the basis of active computation of semantic and/or pragmatic plausibility via inference.

There is evidence which favors the latter interpretation. First, the performance of low-level Markovian POS taggers is actually not very good by the standards of human sentence processing. The average length of the sentences in the Wall Street Journal corpus is around 23.5 words. 97% precision means an error in over half the sentences of average length in this type of written text, and Ratnaparkhi (1998) shows that in practice this is about the proportion for all sentences in the corpus. (Using a 96.3% accurate tagger, he finds only 47.7% of sentences in the Wall Street Journal corpus are error-free). This means something other than n-gram frequency must be doing some of the work in humans. If something else is doing some of the work, then it may be doing all of it.

There is also experimental evidence in work by Steven Crain, Gerry Altmann, and the present author, which shows that parsing decisions are sensitive to the relative plausibility of the semantic interpretations of rival analyses, which may depend on quite transient and rapidly-changing properties of the referential context. This fact has implications even for sentences presented in isolation (see Crain 1980; Crain & Steedman 1985; Altmann 1988; Altmann & Steedman

1988, and Steedman & Altmann 1989 for discussion).

They showed for example that the mention of a single policeman or a set of policemen in the discourse immediately preceding examples like (2a) can effect the tendency to assume that the verb is indicative rather than participial. The authors argued that this was because the presence of multiple individuals or tokens of a given type in a hearer's mental model of the situation under discussion makes the use of a restrictive relative clause or other modifier pragmatically felicitous, whereas the presence of a single individual of a given type makes it redundant and infelicitous. Moreover, they argued that in the null context where no policeman at all has been mentioned, so that a referent and attendant presuppositions must be "accommodated" or added to the contextual model, the accommodation of one individual is less effortful than the accommodation of several together with further distinguishing properties presupposed by the modifier.

If preferences for parsing decisions can be reversed by a few preceding context-setting sentences, then it becomes implausible to argue that the human parser's decision is made on the basis of global frequencies collected over large volumes of input. While finite state approximation techniques can be quite immediately generalized to sequences of words and categories that cross sentence boundaries, as Seidenberg & MacDonald propose in this issue, it is dangerous to assume that this will capture transient referential effects of the kind established by these experiments. Since definite expressions can refer to entities that are merely inferable from the things that have actually been mentioned, there seem to be simply too many ways of getting sets of policemen into the context for it to be possible to collect appropriate statistics based on words and word senses alone. The only alternative seems to be to assume that interpretations are incrementally computed for rival analyses, which are then compared, leading to rapid elimination of less plausible alternatives. But it it is hard to believe that this can be done without a fuller grammatical analysis than that implicit in the SRN.

The study of syntactic priming (Bock 1986) in the paper by Dell et al. in this issue might appear at first glance to encourage a more optimistic view. It uses a variant of the simply recurrent architecture linking the context units of an SRN trained to associate words in sequence with content vectors (or rather, a simulation using a transition network of the successive states of such context units) to those of a production network producing sequences from content. Crucially, pairs of sequences corresponding to active vs. passive surface forms are associated with the same content. Having trained the network, the authors are able to show that a further presentation of a string in (say) active voice biases immediately subsequent productions towards that pattern, even for different content, with results comparable to Bock's subjects. As the authors claim, this shows that apparently syntactic priming effects can emerge from implicit learning rather than from rule-activation or structures in short-term memory.

However, this observation is entirely compatible with the idea that the nature of this implicit learning lies in a change to the probabilities in an implicit finite state machine. As we have seen, it does not follow that the rules themselves are implicit, or that interpretation can be done on the basis of the probabilities alone. It merely shows that "syntactic priming" is not necessarily syntactic at all. Nor does it follow that the analogous priming effects in humans are mediated by actual probabilities, because of the confounding of probability with actively-computed semantics and pragmatic inference discussed in connection with Tabor and Tanenhaus, and Crain's and Altmann's experiments.

The relation of recurrent networks to finite state machines of a more traditional sort such as POS taggers and HMMs (which is of course evident in Jordan's elegant application of recurrent networks to model coarticulation in speech and other motor control problems) suggests a further direction in which connectionist models of syntactic processing might evolve. The trend in symbolic stochastic language processing is away from grammar-independent POS tagging, and towards a greater integration of probabilistic information with the grammar and recursive parsing algorithms, and in particular with the lexicon. This requires a rather different kind of mechanism.

4 Recursive Auto-Associative Memory and Grammar

A number of connectionist processors have used nets as distributed representations of structure, and such networks can be viewed as encoding the thematic roles of propositions. Early versions of the idea such as McClelland & Kawamoto (1986) were non-recursive, but Pollack (1990) showed how recursively embedded structure could be built in such rule-like nets, in an architecture called the Recursive Auto-Associative Memory (RAAM). This was a more efficient version of an even simpler device called the Associative Net (Willshaw, Buneman, & Longuet-Higgins, 1969; Willshaw 1981). An associative net acts as a distributed memory associating pairs of input and output vectors, as in Figure 2, which represents a grid of horizontal input lines and vertical output lines with binary switches (triangles) at the intersections. To store an association

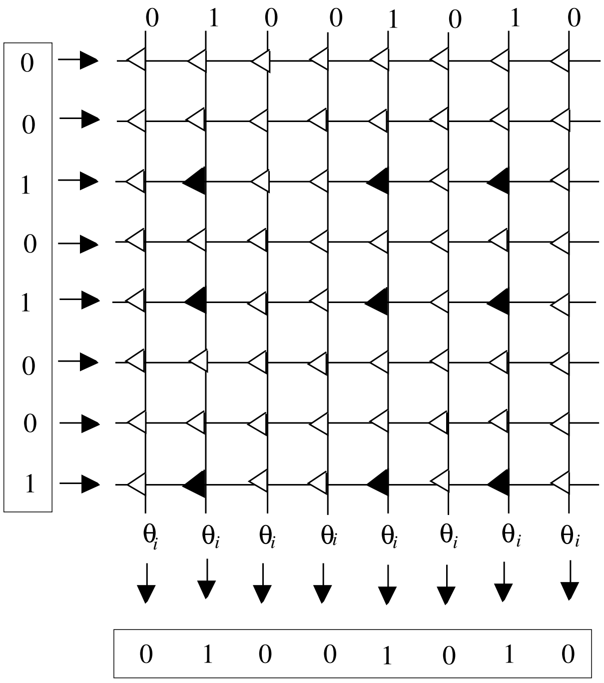

Figure 2
An Associative Net Storing a Single Pointer

between the input vector on the left and the output vector along the top, switches are turned on (black triangles) at the intersection of lines which correspond to a 1 in both input and output patterns. To retrieve the associate of the input, a signal is sent down each horizontal line corresponding to a 1 in the input. When such an input signal encounters an “on” switch, it increments the signal on the corresponding output line by one unit. These lines are then thresholded at a level corresponding to the number of on-bits in the input. With such thresholding, an associative memory can store a number of associations in a distributed fashion, with interesting properties of noise- and damage- resistance. The point of the device for present purposes is that the association of an input vector with an output vector can be regarded as analogous to storing one or more pointers between addresses. Since the output can in turn be used as an input and associated with a further output, an associative memory can be used to store recursive structures of any depth, subject only to information-theoretic limits dependent upon size. Smolensky’s 1990 tensor product representation is a generalization of the same idea.

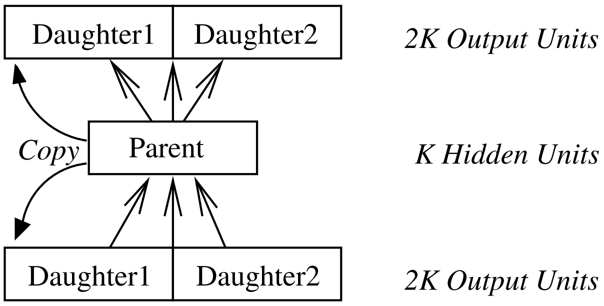

Figure 3
Architecture of a Binary Recursive Auto-Associative Memory (RAAM)

The RAAM mainly differs from the primitive associative net in using hidden units rather than observable switches to encode the association more efficiently. This is achieved by realizing the device as a three-level feed-forward network with the input and output units structured into n sectors each large enough to copy the hidden units into, where n is the maximum branching factor of the nodes in the structure. (Alternatively, as in the case of the associative net, we could regard each pointer as the responsibility of n separate associative devices. We shall see that we do not actually need the ordering information implicit in the standard RAAM). A binary version of the device is sketched in Figure 3.

A recursive structure can be stored bottom-up in the RAAM starting with the leaf elements by recursively auto-associating vectors comprising up to n hidden unit activation patterns that resulted from encoding their daughters. The activation pattern that results from each auto-association of the sets of daughters can then be treated as the address of the parent. Since by including finitely many further units on the input and output layers we can associate node-label or content information with the nodes (a variant that is sometimes referred to as a “Labelling” RAAM, or LRAAM), this device can store recursive parse structures, thematic representations, or other varieties of logical form of sentences.

The device should not be confused with a parser: it is trained with fully articulated structures which it merely efficiently stores. However, the hidden units can be regarded as encoding to some approximation the context-free productions that defined those structures, in a fashion similar to the way Hinton (1990b) encoded part-whole relations. In that sense the device can be held to be capable of inducing the corresponding grammar from the trees (Pollack, 1990, p.88-9). It is also the basis for a very limited degree of generalization (p.94, also of a corresponding tendency to decode novel trees as members of the training set). In theory this generalization could be recursively productive, but in practice this does not seem to have been achieved, possibly because of sever practical limitations on RAAM, which is slow to train, and inherits poor scaling properties from its use of back-propagation. Associative devices more closely related to Willshaw nets, such as the Holographic Reduced Representations (HRR) proposed by Plate (1991, 1994, 1997) are a promising alternative for both recursive structure building and grammar representation.

Pollack proposed to augment the trained RAAM with a variant of the recursive network parsers discussed in the last section to make a parser-interpreter, along lines suggested by St. John & McClelland (1990) for its non-recursive predecessor, and there have been many related proposals since, including Berg (1992); Sharkey & Sharkey (1992); Mikkulainen (1993, 1995). The earlier remarks about limitations on the sense in which SRNs can be said to represent grammars, as opposed to more primitive notions like current state of the FSM or corresponding part-of-speech, show up in the extent to which such devices generalize to truly novel sentence structures. Some of the above authors augment the SRN/RAAM with further modules, but to the extent that these duplicate the standard finite state control and stack architecture of the symbolic approach, they lack both the simple virtues of the RAAM and the wild romance of the pure SRN.

This suggests that it might be better to reserve associative devices like RAAM and HRR for inducing the grammar, in the following sense. Linguists and other symbolists often think of grammar induction as the problem of inducing structure from strings, a problem to which without some external source of information an exact solution is impossible for all interesting cases (Gold, 1967). While for interesting classes of grammar the task can technically be approximated to any degree of exactness, as is shown by the work considered here, the amounts of data and the computational resources that are required for realistically sized cases are psychologically quite daunting. This has led to symbolist claims of innate linguistic knowledge which any red-blooded empiricist is bound to bridle at. However, the only plausible source for such pre-grammatical knowledge has always been semantics, under the assumption that the child comes to language learning equipped with universal conceptual structures on which language-specific grammar is rather directly hung by pairing words and sentences with conceptual structures describing the situation of utterance. (See Chomsky 1965, although Chomsky has always insisted that our access to the detailed nature of conceptual structures, other than via syntax itself, is so inadequate as to make the observation empirically useless).

Gleitman (1990) and Fisher, Hall, Rakowitz, & Gleitman (1994) have rightly pointed to the dangers of identifying "situation of utterance" with "instantaneous state of the physical world," and warned against the assumption that the situation uniquely identifies the relevant conceptual structure or proposition. But Siskind (1996) has shown that under more reasonable assumptions about the nature of the mental representations and the nature of the ambiguity, the information needed for verb learning is available.

The relevance of this point for the present discussion is that conceptual structures can reasonably be regarded as the structural input to a device such as the RAAM for purposes of induction of the underlying grammar. Of course, this is not the whole of language acquisition, because conceptual structure represents universal grammar rather than any specific language—indeed, conceptual structures should properly be regarded as unordered. However, if we regard the RAAM or related device as storing word-meanings as logical forms, rather than sentences, then we can pair those logical forms with language-specific categories. In any of the lexicalist frameworks discussed in the next section, such categories amount to a specification of most if not all of the language-specific grammar.

Such categories could provide the input to a standard-architecture modular symbolist parser using RAAM, HRR, or some other associative device to build interpretable structure, the distinctively connectionist contribution lying in the distributed lexical entries and logical form. The interesting point of such a representation is that we might assume that during training conceptual structures are available prelinguistically, and result relatively directly from the structure of connections to the sensorium, short term memory, and the like. Much of this structure is undoubtedly the result of biological evolution over a very long period, as Wilkins & Wakefield (1995) have pointed out. At higher levels, such structure may arise from non-linguistic network concept-learning of the kind discussed by Hinton (1990b), without mediating symbolic forms.

5 Networks and the Lexicon

The representations that would be built by the RAAM or related distributed associative memory under these assumptions would embody the traditional local domain defined by lexical entities like verbs and their arguments including the subject. There is a growing consensus across linguistic theories that the lexicon is the main locus of language-specific grammatical information, and that what we might loosely call "heads" are lexically specified as controlling such a local domain, as in Lexical-Functional Grammar (LFG, see Bresnan 1982), Combinatory Categorial Grammar (CCG, see Steedman 1985, 1996), Head-Driven Phrase-Structure Grammar (HPSG, see Pollard & Sag 1987, 1994), Lexical Tree-Adjoining Grammar (LTAG, see Joshi & Schabes 1992), and certain versions of the Government-Binding theory (GB, see e.g. Hale & Keyser 1993 and Grimshaw 1997).

The advantage of such theories lies in a closer integration of the lexicon, syntax, semantics, and phonology including intonation, as called for by Kelly (1992); Kelly & Martin (1994) and by Christiansen, Allen, and Seidenberg 1998. For example, in CCG, each word and constituent is associated with a directional syntactic type, a logical form, and a phonological type. "Combinatory" syntactic rules combine such entities to produce not only standard constituents associated with the same three components, such as predicates or VPs, but also non-standard constituents corresponding to substrings such as *I have found*. The latter are involved in phenomena such as coordination and intonational phrasing, as in (3) and (4) (in which % marks an intonational phrase boundary marked by a rise and/or lengthening, and capitals indicate pitch accent or stress).

(3) You have lost, and I have found, a quarter.

(4) Q: I know you lost a DIME, but what have you FOUND?
A: I have FOUND% a QUARTER

The intonational phrasing in the latter example is related in Steedman (1991) to discourse-information structural notions like theme/topic, rheme/comment, and focus/new information. Such non-standard constituents are also claimed by Crain & Steedman (1985) and Altmann & Steedman (1988) to provide direct grammatical support for the fine-grain incremental interpretation by the processor that is implicated by Crain's and Altmann's experimental results.

Within such frameworks, grammar acquisition reduces to decisions such as whether the syntactic type corresponding to (say) the *walking* concept looks for its subject to the left or to the right in the particular language that the child is faced with—in CCG terms, whether it is $S\backslash NP$ or S/NP. Since directionality can be represented as a value on an input unit, and since the categories themselves can be defined as finite state machines, this can be handled by network lexicon learning using devices like RAAM, SRN and the like. Part of the interest of this proposal lies in the possibility that such learning might capture word-order generalizations over the lexical categories, a point that is made by Christiansen & Devlin (1997). It is also important to note once more that any assumption of a covering grammar, not transparent to semantics in this way, such as the Finite State Machine implicit in the SRN, complicates the task of associating meanings with categories very greatly, and appears to pose equivalent difficulties for any attempt to explain the evolution of the language faculty.

These observations suggest that there might be a closer relation between the connectionist and symbolist theories than is usually assumed. Grimshaw in particular relates the forms that categories can take to a constraint-satisfaction problem that can be elegantly solved within Optimality Theory (Prince & Smolensky, 1997) a branch of the Neural Network literature discussed by Smolensky in this issue. Since this process of ordered constraint satisfaction is best seen as a definition of the notion "possible human lexicon," rather than as a process that the parser goes through, the connection is likely to be at the level of language acquisition and machine learning, rather than processing as such, as in the work discussed by Seidenberg 1997 and Seidenberg & MacDonald here and elsewhere. Constraints such as that semantically related categories (such as tensed transitive verbs) tend to have the same directionality in a given language (such as the English SVO order) are "soft" constraints, which can have exceptions (such as English auxiliary verbs)—one of the main motivations behind Optimality Theory. Since Optimality-Theoretic constraint systems can be regarded (under some assumptions at least) as defining Finite State Transducers (Eisner, 1997), it seems likely that the associative memory -based lexical acquisition device sketched above might be suited to acquiring such soft-constraint-based lexicons, as an interesting alternative to the learning mechanisms proposed by Tesar (1997). If so, then the claim that the form of possible lexicons was "emergent" from the neural mechanism would have some force.

A similar tendency towards lexical involvement is evident in current statistical computational linguistic research as well. Recent proposals by Collins (1997) and Charniak (1997) move away from autonomous Markovian POS tagging and prefiltering, and towards a greater integration of probabilities with grammar. Collins in particular uses a standard dynamic-programming-based parsing algorithm under the guidance of probabilities based on dependencies between heads—such as those between main verbs and the nouns that head their arguments. This architecture is quite directly compatible with lexicalized grammars such as CCG, HPSG and TAG. It would be interesting to investigate the relation between Collins' procedure for inducing these probabilities and the neural mechanisms discussed here, which again seem well-suited to capture such dependencies.

6 Semantics and Neural Networks

To observe that both resolution of syntactic non-determinism by human parsers and language acquisition appear to depend upon a structurally explicit semantics manipulated by stack automata might appear to be a sort of underhand appeal for capitulation to the symbolist view. Instead, I want to argue that the most important contribution of sub-symbolic theories to the problem of language understanding may be at the level of the semantics, rather than syntax, for the following reason.

The conclusion that the decisions of the human sentence processor depend on semantics is quite depressing for those of us who need to build practical computational parsers, because we know that the semantics in question is very poorly understood, and that there are no knowledge representation systems that can support it affordably for other than small restricted domains. For that reason we can expect applied computational linguists to keep on using statistics instead, despite the fact that we can be pretty sure that this tactic will never be entirely successful.

However, there is no reason for connectionists to be inhibited in this way, and there is a good reason for them to concentrate their attention elsewhere. The reason our grasp of semantics is so inadequate is undoubtedly that the conceptual primitives that underlie language are grounded in very obscure ways in our physical, social, and intellectual interactions with the real world. It is likely that in many cases the forms they take are directly related to the physical structure of the sensory-motor apparatus. In most cases it seems likely that, as Fodor (1975) has claimed, there really is very little decomposition of meaning below the level of the morpheme. That is, even if one believes (*pace* Fodor, whose arguments to the contrary do not really apply to lexicalized grammars of the kind assumed here) that the logical form of the verb *kill* involves the composition of a CAUSE predicate and a DIE predicate, that is about as far as it goes. The CAUSE primitive doesn't seem to want to decompose any further, and in fact shows signs of being distinct from the translation of the word *cause*. In fact all one seems to be able to do is to define the inference system directly in terms of meaning postulates relating these morpheme- or near-morpheme- level primitives directly, as Fodor proposed. This non-decomposability of lexical meaning shows itself in may ways, from the nonexistence of a concept that means exactly the same thing as *waterproof* but applies to non-physical entities such as integers, to the fuzziness of the verb classification schemata of Levin (1993) discussed by Dang, Rosenzweig, & Palmer (1997). The latter looks more like the result of a principal components or factor analysis than a semantically interpretable set of features, despite its strong syntactic foundation.

This again seems to be exactly the kind of system that the sub-symbolic approach is best-suited to analyze. If the above remarks are correct then we would not expect a principal components analysis to be interpretable in the way that we expect the results of parsing to be, and would be happy to tolerate a high degree of cognitive impenetrability in return for the efficiency and learnability of distributed representation. This again would really be emergence worthy of the name.

7 Conclusion: Project for a Scientific Psycholinguistics

A project of the kind outlined above for the development of a grounded semantics will require starting at a much earlier stage than the onset of language learning. It is likely that it will have to recapitulate in neural computational terms the kind of program of sensory-motor development outlined in Piaget 1952 (though much theoretical baggage can be dispensed with, particularly in the light of more recent research on the status of "preoperational" and "formal operational" components). Work along these lines has already been sketched in more symbolic computational terms by Drescher (1991) and Siskind (1995).

It is likely that such a research program would proceed by first conceptualizing primary bodily actions and sensations, then coordinating perception and primary actions like reaching, then conceptualizing identity, permanence and location of objects, first independent of their percepts, then of the particular actions they are involved in, amounting to the internalization of the components of a stable world independent of the child's actions. Later stages would have to include the conceptualization of more complex events including intrinsic actions of objects themselves (such as falling), translations and events involving multiple participants, intermediate participants including tools, and goals. At this final stage of purely sensory-motor development most of the prerequisites for language learning would be established, perhaps embedded in RAAM or some other associative memory, and could be used to support a program of inducing a similarly layered sequence of linguistic categories such as: deictic terms based on a proximal/distal dimension (whose central place in language development with respect to reference and definiteness is discussed by Lyons 1977—cf. Freud 1920, pp.11-16 for a revealing case study), markers of topic, comment and contrast, common nouns, spatial and path terms, causal verbs, modal and propositional attitude verbs, and temporal terms. It is likely that the semantic theory that would emerge from this work would be rather unlike anything proposed so far within standard logicist frameworks. Such a semantics would be likely to make us view phenomena like quantification, modality, negation, and variable-binding in new ways, within a unified theory combining symbolic and neurally-grounded levels.

It is probably too soon to tell whether the distributed computational devices of the connectionist approach make such an ambitious research program any more feasible than it was at the time of Freud's unpublished *Project for a Scientific Psychology*, arguably the first manifesto for a cognitive neuroscience (albeit a localist one), which he abandoned forever in 1895 in favor of the symbolic approach. The success of the present project is likely to depend crucially on the involvement of more reliable and biologically plausible network models than the three simple types of feed-forward networks considered above. (In particular, in order to concentrate on the general relation of this class of models to symbolist alternatives, I have only referred in passing to some well-known problems with techniques specifically based on back-propagation, which include poor scaling of data-set size and learning times as the number of connections increases. Hinton & Gharamani 1997 presents an interesting recent alternative.) Nevertheless, the papers collected here make a convincing case that the attempt must be made, and has already begun.

References

Allen, J. (1997). Probabilistic constraints in acquisition. In *GALA 3: Language Acquisition: Knowledge Representation and Processing*, pp. 300–305. Edinburgh University.

Altmann, G. (1988). Ambiguity, parsing strategies, and computational models. *Language and Cognitive Processes*, *3*, 73–98.

Altmann, G. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191–238.

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, Cambridge MA. MIT Press.

Black, A. (1989). Finite state machines from feature grammars. In *Proceedings of the 1989 International Parsing Technologies Workshop, Pittsburgh, August*, pp. 277–285.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.

Bresnan, J. e. (1982). *The mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press.

Bresnan, J. W., Kaplan, R. M., Peters, S., & Zaenen, A. (1982). Cross-serial dependencies in dutch. *Linguistic Inquiry*, *13*, 613–636.

Bridle, J. (1992). Neural networks or hidden markov models for automatic speech recognition: Is there a choice? In P. Laface & R. De Mori (eds.), *Speech Recognition and Understanding: Recent Advances, Trends and Applications*, no. 75 in NATO ASI Series: Advanced Science Institutes Series F: Computer and Systems Sciences, pp. 225–236. Springer.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proceedings of the 3rd Conference on Applied Computational Linguistics*, pp. 152–155, Trento.

Casey, M. (1996). The dynamics of discrete-time computation, with application to recurrent neural networks and finite state machine extraction. *Neural Computation*, *8*, 1135–1178.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference of the American Association for Artificial Intelligence, Providence, RI., July*, pp. 598–603.

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

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge MA: MIT Press.

Christiansen, M., Allen, J., & Seidenberg, M. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221–268.

Christiansen, M. & Chater, N. (1994). Generalization and connectionist language learning. *Mind and Language*, *9*, 273–287.

Christiansen, M. & Chater, N. (1998). Towards a connectionist model of recursion in human linguistic performance. *Cognitive Science*. in press.

Christiansen, M. & Devlin, J. (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *Proceedings of the 19th Annual Cognitive Science Society Conference*, pp. 113–118, Mahwah, NJ. Lawrence Erlbaum Associates.

Church, K. (1988). A stochastic parts program and a noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing, Austin TX, February*, pp. 136–143. Association for Computational Linguistics, MIT Press.

Cleeremans, A. (1993). *Mechanisms of Implicit Learning*. Cambridge MA: MIT Press.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. (1995). Graded state machines: The representation of temporal contingencies in feedback. In Y. Chauvin & D. E. Rumelhart (eds.), *Backpropagation : Theory, Architectures, and Applications*, Developments in connectionist theory, pp. 274–312. Hillsdale, N.J: Lawrence Erlbaum Associates.

Collins, M. (1997). Three generative lexicalized models for statistical parsing. In *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics, Madrid, July*, pp. 16–23.

Crain, S. (1980). Pragmatic constraints on sentence comprehension. Ph.D. thesis, University of California at Irvine.

Crain, S. & Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological parser. In L. K. D. Dowty & A. Zwicky (eds.), *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, ACL Studies in Natural Language Processing, pp. 320–358. Cambridge University Press.

Dang, H. T., Rosenzweig, J., & Palmer, M. (1997). Associating semantic components with levin classes. In *Proceedings of Interlinga Workshop, MTSUMMIT97, San Diego,CA, Oct 28*, pp. 11–18.

Dell, G., Chang, F., & Griffin, Z. (1998). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*. this issue.

Drescher, G. (1991). *Made-up Minds*. Cambridge, MA.: MIT Press.

Eisner, J. (1997). Efficient generation in primitive optimality theory. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Association for Computational Linguistics, Madrid, July*.

Elman, J. (1990). Representation and structure in connectionist models. In G. Altmann (ed.), *Cognitive Models of Speech Processing*, pp. 345–382. Cambridge MA: MIT Press.

Elman, J. (1995). Language as a dynamical system. In R. Port & T. van Gelder (eds.), *Mind as Motion*, pp. 195–225. Cambridge, MA: MIT Press.

Fisher, C., Hall, G., Rakowitz, S., & Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, *92*, 333–375.

Fodor, J. (1975). *The Language of Thought*. Cambridge MA: Harvard.

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *35*, 183–204.

Frazier, L. & Clifton, C. (1996). *Construal*. Cambridge MA: MIT Press.

Freud, S. (1895). A project for a scientific psychology. In *The Origins of Psychoanalysis: Letters to William Fliess*. London: Imago. Pub. 1954, Standard Edition, I, 295-343.

Freud, S. (1920). *Beyond the Pleasure Principle*. London: Hogarth Press. Standard Edition, XVIII, 7-64.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 1–55.

Gold, E. (1967). Language identification in the limit. *Information and Control*, *16*, 447–474.

Grimshaw, J. (1997). Projection, heads, and optimality. *Linguistic Inquiry*, *28*, 373–422.

Hadley, R. (1994a). Systematicity in connectionist language learning. *Mind and Language*, *9*, 247–272.

Hadley, R. (1994b). Systematicity revised: Reply to christiansen and chater and niklasson and van gelder. *Mind and Language*, *9*, 431–444.

Hale, K. & Keyser, S. J. (1993). On argument structure and the lexical expression of syntactic relations. In K. Hale & S. J. Keyser (eds.), *The View from Building 20*, pp. 53–109. Cambridge MA: MIT Press.

Hinton, G. (ed.) (1990a). *Connectionist Symbol Processing*. Cambridge Mass.: MIT Press/Elsevier. Reprint of *Artificial Intelligence, 46:1-2*.

Hinton, G. (1990b). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, *46*, 47–75. Reprinted in Hinton (1990a).

Hinton, G. (1990c). Preface to the special issue on connectionist symbol processing. *Artificial Intelligence*, *46*, 1–4. Reprinted in Hinton (1990a).

Hinton, G. & Gharamani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London, B*, *352*, 1177–1190.

Huybregts, R. (1984). The weak inadequacy of context-free phrase-structure grammars. In G. de Haan & a. W. Z. Mieke Trommelen (eds.), *Van Periferie naar Kern*. Foris Dordrecht.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the Institute for Electronic and Electrical Engineers*, *64*, 532–556.

Jordan, M. (1989). Serial order: a parallel distributed processing approach. In J. Elman & D. Rumelhart (eds.), *Advances in Connectionist Theory*. Hillsdale, NJ: Erlbaum.

Joshi, A. & Schabes, Y. (1992). Tree adjoining grammars and lexicalized grammars. In M. Nivat & M. Podelski (eds.), *Definability and Recognizability of Sets of Trees*. Elsevier, Princeton.

Kelly, M. (1992). Using sound to solve syntactic problems. *Psychological review*, *99*, 349–364.

Kelly, M. & Martin, S. (1994). Domain-general abilities applied to domain-specific tasks: Sensitivity to probabilities in perception, cognition, and language. *Lingua*, *92*, 105–140.

Kim, A., Srinivas, B., & Trueswell, J. (1998). The convergence of lexicalist perspectives in psycholinguistics and computational linguistics. In P. Merlo & S. Stevenson (eds.), *Papers from the Special Section on the Lexicalist Basis of Syntactic Processing, CUNY Conference, Rutgers, March*. Amsterdam: John Benjamins.

Landauer, T. & Dumais, S. (1997). A solution to plato's problem: the latent semantic analysis of the acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T., Laham, D., Rehder, B., & Schreiner, M. (1997). How well can passage meaning be derived without using word order? In *Proceedings of the Nineteenth Conference of the Cognitive Science Society*.

Levin, B. (1993). *English Verb-Classes and Alternations: A preliminary Study*. Chicago, IL.: University of Chicago Press.

Lyons, J. (1977). *Semantics, Vol.II*. Cambridge: Cambridge University Press.

McClelland, J. & Kawamoto, A. (1986). Mechanisms of sentence processing: Assigning roles to constituents. In J. McClelland, D. Rumelhart, & the PDP REesearch Group (eds.), *Parallel Distributed Processing*, pp. 272–325. Cambridge MA: MIT Press.

McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1997). Modelling the influence of thematic fit (and other constraints) in online sentence comprehension. *Journal of Memory and Language*, *in press*.

Merialdo, B. (1994). Tagging english text with a probablistic model. *Computational Linguistics*, *20*, 155–171.

Mikkulainen, R. (1993). *Subsymbolic Natural Language Processing*. Cambridge MA: MIT Press.

Mikkulainen, R. (1995). Subsymbolic parsing of embedded structures. In R. Sun & L. Bookman (eds.), *Computational Architectures Integrating Neural and Symbolic Processes*, pp. 153–186. Dordrecht: Kluwer.

Montague, R. (1970). English as a formal language. In B. Visentini (ed.), *Linguaggi nella Società e nella Technica*, pp. 189–224. Milan: Edizioni di Comunità. Reprinted in Montague 1974, 188-221.

Montague, R. (1974). *Formal philosophy: papers of Richard Montague*. New Haven: Yale University Press. Edited by Richmond H. Thomason.

Niklasson, L. & van Gelder, T. (1994). On being systematically connectionist. *Mind and Language*, *9*, 288–302.

Pearlmutter, B. (1995). Gradient calculations for dynamic recurrent neural networks: a survey. *IEEE Transactions on Neural Networks*, *6*, 1212–1228.

Piaget, J. (1952). *The Origins of Intelligence in Children*. New York: Norton.

Plate, T. (1991). Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence, San Mateo CA*, pp. 30–35, Los Altos CA. Morgan Kaufmann.

Plate, T. (1994). Distributed representations and nested compositional structure. Ph.D. thesis, University of Toronto.

Plate, T. (1997). Structure matching and transformation with distributed representations. In R. Sun & F. Alexandre (eds.), *Connectionist-Symbolic Integration*. Hillsdale, NJ: Lawrence Erlbaum.

Pollack, J. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*, 77–105. Reprinted in Hinton (1990a).

Pollard, C. & Sag, I. (1987). *Information-based Syntax and Semantics*, vol. 1. Chicago: CSLI/Chicago University Press.

Pollard, C. & Sag, I. (1994). *Head Driven Phrase Structure Grammar*. CSLI/Chicago University Press, Chicago.

Prince, A. & Smolensky, P. (1997). Optimality: from neural networks to universal grammar. *Science*, *275*, 1604–1610.

Ratnaparkhi, A. (1998). Maximum entropy models for natural language ambiguity resolution. Ph.D. thesis, University of Pennsylvania.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing, Vol. I: Foundations*, pp. 318–362. Cambridge MA: MIT Press.

Seidenberg, M. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*, 1599–1603.

Seidenberg, M. & MacDonald, M. (1998). A probabilistic constraints approach to language acquisition and processing constraints. *Cognitive Science*. This issue.

Sharkey, N. & Sharkey, A. (1992). A modular design for connectionist parsing. In A. Nijholt & M. Drossaers (eds.), *Twente Workshop on Language Technology 3: Connectionism and Natural Language Processing*, pp. 87–96. Enschede, the Netherlands: Dept of Computer Science, University of Twente.

Shieber, S. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, *8*, 333–343.

Siskind, J. (1995). Grounding language in perception. *Artificial Intelligence Review*, *8*, 371–391.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*, 159–216. Reprinted in Hinton (1990a).

Smolensky, P. (1998). Grammar-based connectionist approaches to language. *Cognitive Science*. This issue.

Spivey-Knowlton, M., Trueswell, J., & Tanenhaus, M. (1993). Context effects in syntactic ambiguity resolution. *Canadian Journal of Psychology*. incomplete reference.

Srinivas, B. (1997). Complexity of lexical descriptions and its relevance to partial parsing. Ph.D. thesis, University of Pennsylvania. IRCS Report 97-10.

Srinivas, B. & Joshi, A. (1994). Disambiguation of super parts of speech (or supertags): Almost parsing. In *Proceedings of the International Conference on Computational Linguistics (COLING 94), Kyoto University, Japan*.

St. John, M. & McClelland, J. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*, 217–257. Reprinted in Hinton (1990a).

Steedman, M. (1985). Dependency and coordination in the grammar of dutch and english. *Language*, *61*, 523–568.

Steedman, M. (1991). Structure and intonation. *Language*, *67*, 262–296.

Steedman, M. (1996). *Surface Structure and Interpretation*. Cambridge Mass.: MIT Press. Linguistic Inquiry Monograph, 30.

Steedman, M. & Altmann, G. (1989). Ambiguity in context: a reply. *Language and Cognitive Processes*, *4*, 105–122.

Tabor, W. & Tanenhaus, M. (1998). Dynamical models of sentence processing. *Cognitive Science*. This issue.

Tesar, B. (1997). An iterative strategy for language learning. *Lingua*. in press.

Trueswell, J. & Tanenhaus, M. (1992). Consulting temporal context in sentence comprehension: Evidence from the monitoring of eye movements in reading. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, Hillsdale, NJ. Erlbaum.

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*, 285–318.

Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 528 – 553.

Wilkins, W. & Wakefield, J. (1995). Brain evolution neurolinguistic preconditions. *Behavioral and Brain Sciences*, pp. 161–182.

Williams, C. & Hinton, G. (1990). Mean field networks that learn to discriminate temporally distorted strings. In D. Touretsky, J. Elman, T. Sejnowski, & G. Hinton (eds.), *Connectionist Models: Proceedings of the 1990 Summer School*.

Willshaw, D. (1981). Holography, association and induction. In G. Hinton & J. Anderson (eds.), *Parallel Models of Associative Memory*, pp. 83–104. Hillsdale NJ: Erlbaum.

Willshaw, D., Buneman, P., & Longuet-Higgins, C. (1969). Non-holographic associative memory. *Nature*, *222*, 960–962.

Woods, W. (1970). Transition network grammars for natural language analysis. *Communications of the Association for Computing Machinery*, *18*, 264–274.