Massively-Parallel Inferencing for Natural Language Understanding and Memory Retrieval in Structured Spreading-Activation Networks

Trent E. Lange

Artificial Intelligence Laboratory
Computer Science Department
University of California, Los Angeles
Los Angeles, CA 90024

INTRODUCTION

One of the most difficult parts of the natural language understanding process is forming a semantic interpretation of the text. A reader must often make multiple inferences to understand the motives of actors and to causally connect actions that are unrelated on the basis of surface semantics alone. The inference process is complicated by the fact that text is often ambiguous both lexically and pragmatically, and that new context often forces a reinterpretation of the input's meaning. This language understanding process itself does not exist in a vacuum — as people read text or hold conversations, they are often reminded of analogous stories or episodes. The types of memories that are triggered are influenced by context from the inferences and disambiguations of the understanding process. A full model of the language understanding and memory retrieval processes must take into account the interaction of the two and how they effect each other.

An example of some of the inferencing and disambiguation problems of the language understanding process is the sentence *John put the pot inside the dishwasher because the police were coming.* (**Hiding Pot**). In this sentence, it first seems more likely that John is trying to clean a cooking-pot, but later seems (to many people) that he was instead trying to hide some marijuana from the police. This reinterpretation cannot be made without a complex plan/goal analysis of the input being made at some level — an analysis requiring the ability to make multiple inferences from rules about general knowledge.

After a person has read a story, they are sometimes reminded of similar episodes. For example, after reading the **Hiding Pot** sentence, a person might be reminded of an analogous story they had read earlier, such as *Billy put his Playboy under the bed so his mother wouldn't see it and spank him* (**Dirty Magazine**). Of course, a person would likely only be reminded of such a story if they had inferred that John was trying to hide marijuana in the original **Hiding Pot** story. If they had instead stuck with their original interpretation of cleaning a cooking-pot, then they probably wouldn't retrieve the **Mother Hiding** story, and instead might retrieve a story relating to cleaning, if anything at all.

These examples illustrate several of the problems of natural language understanding and memory retrieval. First, people need to be able to make multiple, dynamic inferences very quickly in able to understand language texts. The speed with which they do so indicates that they are probably using a parallel inferencing process to explore many different possible interpretations at once. Secondly, they need to be able to integrate multiple sources of evidence from context to disambiguate words and interpretations, implying that some sort of constraint satisfaction process is used along with the inferencing process to perform disambiguation. And finally, both of these processes seem to effect the kinds of episodes and analogies that people retrieve from memory. This paper provides an overview of ROBIN (Lange & Dyer, 1989) and REMIND (Lange & Wharton, in press), two structured connectionist models that provide a potential explanation for these abilities and that perform natural language understanding and episodic memory retrieval using parallel dynamic inferencing and constraint satisfaction.

PREVIOUS WORK IN SEMANTIC LANGUAGE UNDERSTANDING

Symbolic, rule-based systems have had some success performing the inferencing necessary for language understanding, but have substantial difficulties with resolving ambiguities and performing reinterpretation. On the other hand, while distributed connectionist models using recurrent networks can perform disambiguation and understand sequential, script-based stories (e.g. Miikkulainen & Dyer, 1991, St. John, 1992), it is unclear whether they can be scaled up to understanding language that requires the inference of causal relationships between events for completely novel stories. This requires chains of dynamic inferences over simple known rules, with each inference resulting in a potentially novel intermediate state (Touretzky, 1990). Other distributed connectionist models are able to partially handle this problem by explicitly encoding variables and rules in the network (e.g. Touretzky & Hinton, 1988). Unfortunately, these models are *serial at the knowledge level* — i.e. they can only select and fire one rule at a time, a serious drawback for language un-

derstanding, especially for ambiguous text in which multiple alternative interpretations must often be explored in parallel (Lange & Dyer, 1989).

Marker-passing models (e.g. Hendler, 1988, Norvig, 1989) solve many of these problems by spreading symbolic markers (representing variable bindings) across labeled semantic networks in which concepts are represented by individual nodes. Because of this, they are able to perform dynamic inferencing and pursue multiple candidate interpretations of a story in parallel as markers propagate across different parts of the network. One of the main drawbacks of marker-passing models is the generally all-or-nothing nature of their marker-generated paths. Because of this, they must use a separate symbolic path evaluation mechanism to select the best interpretation path among the often large number of alternative paths generated, a particularly serious problem for ambiguous text. Some efforts have been made at developing hybrid marker-passing models that solve some of these ambiguity problems using constraint satisfaction (cf. Kitano *et al.*, 1989).

Structured spreading-activation networks also have the potential to pursue multiple candidate interpretations of a story in parallel, since each interpretation is represented by activation in different local areas of the network. Unlike marker-passing networks, however, structured spreading-activation models (e.g. Waltz & Pollack, 1985) are ideally suited for disambiguation because it is achieved automatically as related concepts under consideration provide graded activation evidence and feedback to one another in a form of constraint relaxation. The spreading-activation process therefore causes the nodes of the best interpretation in a given context (i.e. the one with the most contextual evidence) to become the most highly-activated, so that the winning interpretation is "chosen" within the network itself. Unfortunately, spreading-activation models have been limited because the activation on their nodes does not tell where it came from (unlike the markers of marker-passing systems), and hence what the variable bindings and inferences of the network are.

LANGUAGE UNDERSTANDING IN ROBIN

Our approach is to develop and explore structured connectionist networks that build upon the advantages of spreading-activation networks and that are capable of supporting the processing abilities necessary for language understanding. To this end, we have developed ROBIN, a structured spreading-activation network that can perform some of the high-level inferencing needed for language understanding (Lange & Dyer, 1989). It does this by having structure within the network that hold uniquely-identifying patterns of activation, called *signatures*, that represent the concepts that are bound to a role. Signatures are propagated across the network (as activation) over paths of binding units, allowing variable binding and parallel inferencing like that of symbolic marker-passing networks. ROBIN's use of signatures for variable binding and inferencing is similar to that of the recent alternative structured connectionist approaches of Shastri & Ajjanagadde (in press) and Sun (in press).

The most important aspect of ROBIN is that the structure allowing signatures to be propagated for inferencing is integrated within a normal semantic spreading-activation network in which individual nodes represent concepts. This allows ROBIN to go beyond the simple rule-firing abilities of most alternative structured and marker-passing binding approaches, because the activation levels of these conceptual nodes represent the amount of *evidence* available for them in the current context. Because of this, ROBIN is able to use constraint satisfaction to disambiguate both lexical and pragmatic ambiguities within the network — the winning interpretation is simply the path of concepts with the highest levels of activation, with their role bindings being given by the signatures on the corresponding binding units.

As an example, see Figure 1, which shows the activations in a simplified portion of the network immediately after input for *"John put the pot inside the dishwasher"* has been presented to the network as clamped activations. The nodes in the lower layer of the network form a normal semantic network whose weighted connections represent world knowledge, e.g. that the action of transferring an object inside of a container (Transfer-Inside) results in it being inside the container (Inside-Of), and that two possible concept refinements (or specialized reasons for) it being inside are because it is inside of a dishwasher (Inside-Of-Dishwasher, which will lead to further unshown inferences about it being cleaned) or because it is inside of an opaque object (Inside-Of-Opaque, which will lead to inferences about it being hidden from sight). In Figure 1, the Transfer-Inside node has been clamped to a high level of activation (darkened oval boundary) to represent the fact that a Transfer-Inside has occurred (from John putting the pot inside the dishwasher).

Signature activations for variable binding and inferencing are held by the black binding units in the top plane of Figure 1. In this simplified example, signatures are arbitrary but uniquely-identifying scalar activation values, e.g. 6.8 stands for Marijuana, 9.2 stands for Cooking-Pot, and 5.4 stands for Cake. As shown in the figure, unit-weighted connections between binding units allow signatures to be propagated to other roles given by the general knowledge rules in the network (Lange & Dyer, 1989). For example, there are connections from the binding units of Transfer-Inside's Object to the respective binding units of Inside-Of's Object, since the object transferred inside is always the object that ends up inside. Similarly, there are connections from the binding units of Transfer-Inside's Location to the respective binding units of Inside-Of's Location (not shown). To represent the fact that Transfer-Inside's Object is known to be either Marijuana or a Cooking-Pot, the binding units of its Object are clamped to the activations of their signatures, 6.8 and

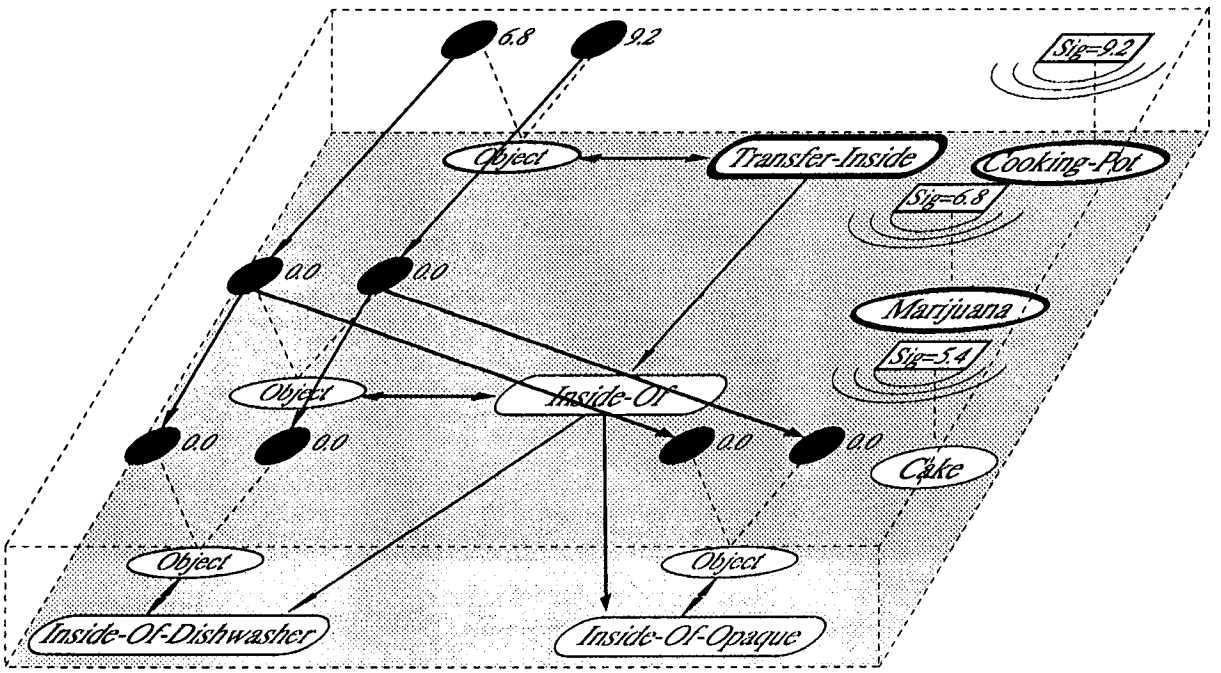

Figure 1. Simplified ROBIN network segment after initial clampings for *"John put the pot inside the dishwasher"*. The figure shows the parallel paths over which evidential activation (bottom plane) and signature activation (top plane) are spread for inferencing. Signature nodes (outlined rectangles) and binding nodes (solid black circles) are in the top plane. Thickness of conceptual node boundaries (ovals) represents their levels of evidential activation. (Node names do not affect the spread of activation in any way. They are simply used to initially set up the network's structure and to aid in analysis.)

9.2, respectively (Figure 1)[1]. Similarly, one of the binding units of Transfer-Inside's Actor role is clamped to the signature of John and of its Location role clamped to the signature of Dishwasher (not shown).

Once the activations of the initial signature bindings and active conceptual nodes of the original phrase are clamped, both types of activation spread through the network. Figure 2 shows the result of this propagation after the network has settled in processing the inputs of Figure 1 for *"John put the pot inside the dishwasher"*. The signature activations representing the bindings have propagated along paths of corresponding binding units, so that the network has inferred that the Cooking-Pot or Marijuana is Inside-Of the Dishwasher (shown by the fact that their signatures are on the appropriate binding units). As can be seen in the figure, the propagation of signatures has also already instantiated two different candidate interpretation paths for the sentence, one going through Inside-Of-Dishwasher (that continues through other cleaning frames) and one going through Inside-Of-Opaque (that continues through frames representing the object being blocked from sight, being hidden, and so forth). Both these interpretations and others are explored in parallel as activation is spread. At the same time, activation spreads and accumulates along the bottom layer of conceptual nodes to disambiguate between these candidate interpretations.

In Figure 2 the Inside-Of-Dishwasher path has ended up with the most activation because of activation feedback between it and its strong stereotypical connections to Cooking-Pot and Dishwasher. The interpretation that the network arrives at for *"John put the pot inside the dishwasher"* is therefore that John was trying to clean a cooking-pot. This interpretation could change, however, if the network was presented with input for *"because the police were coming"*, whose inferences would provide more evidence to hiding and thus the Inside-Of-Opaque path. See (Lange & Dyer, 1989) for details on how ROBIN's networks perform such parallel inferencing and disambiguation.

---

[1]ROBIN does not currently address the problem of deciding upon the original syntactic bindings, e.g. that the meanings of *"pot"* are bound to the Object role of Transfer-Inside. Rather, ROBIN's networks are given these initial bindings (by hand-clamping) and use them for high-level inferencing and disambiguation.

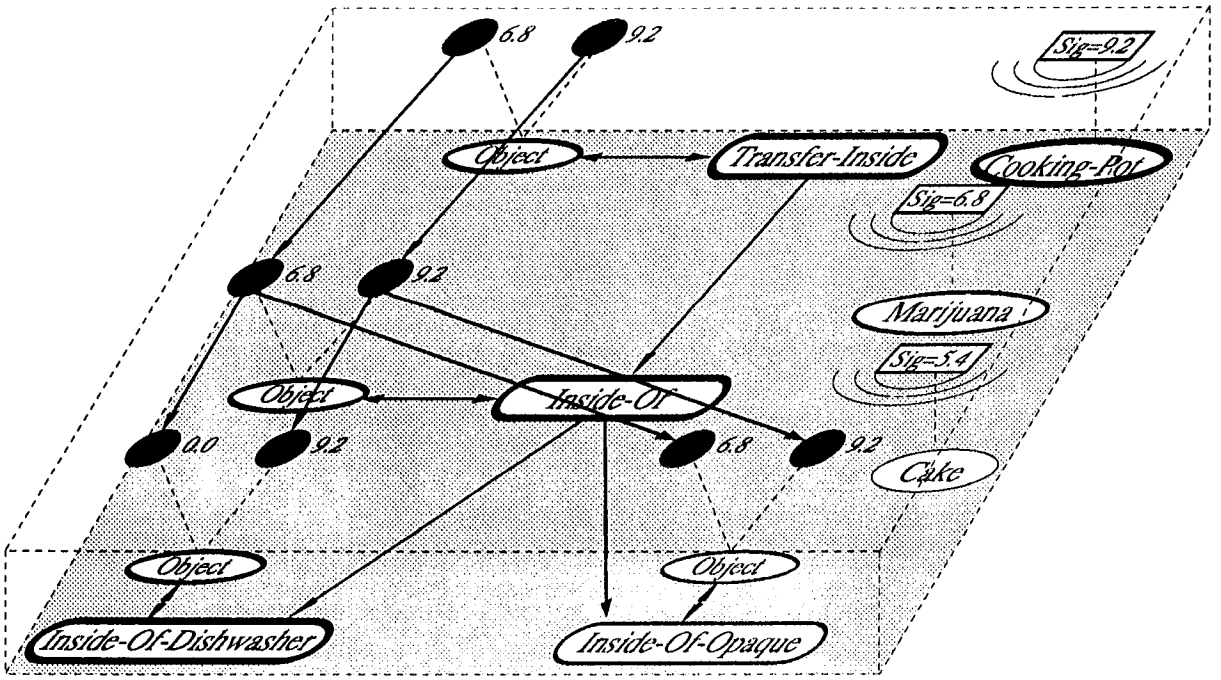

Figure 2. Network of Figure 1 after activation has settled.

MEMORY RETRIEVAL IN REMIND

Using ROBIN as a base, we have also explored how the natural language understanding process can be integrated with memory retrieval. We have built REMIND, a spreading-activation model that performs both inferencing and episodic memory in order to model the effects of inferencing and disambiguation on the retrieval process (Lange & Wharton, in press). While there have been other connectionist models of analogical retrieval (e.g. Barnden & Srinivas, in press, Thagard *et al.*, 1990), few have been built to model and explore the effects of language understanding on retrieval.

REMIND's structured spreading-activation networks encode world knowledge about concepts and general knowledge rules for inferencing in the same way as ROBIN (e.g. Figure 1 and 2). REMIND's networks also contain representations of prior episodes, such as *Fred put his car in the car wash before his date with Wilma* (**Car Wash**) and *Billy put his Playboy under the bed so his mother wouldn't see it and spank him* (**Dirty Magazine**). The representations of these episodes are the actual plan/goal analysis (or discourse model) that was inferred by the network when input for them was first presented to the network to be understood. These prior episodes are indexed into the semantic comprehension network through connections with all the knowledge structures with which they were understood.

To perform retrieval, REMIND is given a short text passage to use as a deliberate memory cue, such as **Hiding Pot.** Units in the network representing the cue and its syntactic bindings are clamped to high levels of activation, which then spreads through the network. By propagating signature activation, the network makes in parallel the different possible inferences that might explain the input, as in ROBIN. Because the units representing long-term memory episodes are connected within the network, an important side-effect of the understanding process is that the episodes having concepts related to the elaborated cue also become highly-activated. This includes episodes related because their is superficial semantic overlap with the cue (e.g., episodes involving other kitchen appliances or illegal drugs) and episodes related abstractly to the cue because they share similar inferred plans and goals (e.g. the **Dirty Magazine** episode becomes activated after understanding **Hiding Pot** because both share the inferences that a person was trying to Hide something to avoid Punishment). After the network settles, the episode that received the most activation from the cue's interpretation and surrounding context becomes the most highly activated, and is therefore retrieved as the best match from the cue (e.g. **Dirty Magazine** is retrieved as the best analogy for **Hiding Pot**).

REMIND is thus an integrated model in which a single mechanism drives both the language understanding and memory retrieval processes. The same spreading-activation mechanism that infers a single coherent interpretation of a cue also activates the episodes the model retrieves from memory. Activation of these episodes combines evidence in a form of constraint satisfaction from both the surface semantics of the input (i.e., different possible word and phrase meanings) and the deeper thematic inferences made from the input, so that the recalled episodes depend on both surface and analogical similarities with the cue. Because of this, REMIND can potentially account for many psychological phenomena involv-

ing priming and language effects in human memory retrieval, such as increased retrieval due to repetition, recency, and prior semantic priming, all of which can be modeled by variations in evidential activation levels prior to presentation of the cue and due to the retrievals themselves (Lange & Wharton, in press).

REMIND's use of signatures to perform massively-parallel inferencing in a structured spreading-activation network as part of the memory retrieval process also gives it several purely computational advantages over previous memory retrieval models. Previous psychological models of analogical retrieval, such as ARCS (Thagard *et al.*, 1990), generally use a costly serial search mechanism to make contact with potential targets to retrieve. Most AI case-based reasoning models (cf. Riesbeck & Schank, 1989) have a similar mechanism to make contact with potential cases, and/or use a serial comparison mechanism to evaluate the cases retrieved and whether or not they contain the desired *indices* for retrieval. It is not clear whether the serial mechanisms that such analogical and case-based retrieval models can operate with sufficient speed when scaled up to the huge size of human memory. REMIND, in contrast, uses its massively-parallel signature inferencing to generate potentially important retrieval indices concurrently, while using the network's constraint satisfaction to narrow down and select the most important of those retrieval indices in a given context.

IMPLEMENTATION ON THE CONNECTION MACHINE

Although serial simulations of connectionist networks can reasonably handle networks of up to medium size, the massively-parallel nature of connectionist networks such as ROBIN and REMIND dictates that large models be simulated on similarly massively-parallel machines. We have therefore implemented our general-purpose object-oriented connectionist simulator, DESCARTES (Lange *et al.*, 1989) on the Connection Machine CM-2, a SIMD machine having up to 64 K processors (Hillis, 1985). Using a generalization of Belloch & Rosenberg (1987)'s algorithm for simulating backpropagation networks with arbitrary connectivity on the CM, the CM version of DESCARTES is able to efficiently simulate arbitrary large-scale heterogeneous neural networks (Lange, 1990). Using the CM version of DESCARTES, we have been able to simulate ROBIN and REMIND networks of over 100,000 units several hundred times faster than the serial simulator's implementation, making research into such large-scale massively-parallel networks feasible.

DISCUSSION

Altogether, ROBIN's structured spreading-activation networks appear to be a promising approach to the problem of inferencing and disambiguation for language understanding. They allow massively-parallel inferencing at the knowledge level in concert with disambiguation and reinterpretation by contextual constraint satisfaction, all within the structure of the network. This automatic disambiguation ability is ROBIN's primary advantage over most symbolic marker-passing systems, which can also generate alternative inference paths in parallel, but which must use a serial path evaluator separate from the marker-spreading process to select the best interpretation, a significant problem as the size of the networks increase and the number of generated inference paths to be evaluated increases dramatically.

Once a connectionist model is able to perform some of the inferencing and disambiguation of the natural language understanding process, it is a natural extension to have the resulting interpretations directly influence memory retrieval, as appears to be the case in people. REMIND is an initial exploration of how this can be done by integrating the networks used for the understanding process (from ROBIN) with subnetworks that represent memory episodes. Because of this, REMIND is a first step towards a model that can potentially account for many psychological phenomena involving priming and language effects in human memory retrieval.

FUTURE WORK

In the future, we plan to concentrate our research in two directions: (1) extending ROBIN to handle more complex semantic language understanding, and (2) modifying the integrated ROBIN/SAARCS model so that it is able to also memorize the stories it understands. In regards to extending ROBIN's language understanding and inferencing abilities, we are currently testing solutions that allow it to handle more than one dynamic instance of the same concept at a time (to enable it to form interpretations for longer input stories) and to handle some of the more complex inference rules (such as recursive inferences) necessary for language understanding, while retaining its parallelism.

REFERENCES

Barnden, J., & Srinivas, K. (in press). Overcoming rule-based rigidity and connectionist limitations through massively-parallel case-based reasoning. *International Journal of Man-Machine Studies.*

Hendler, J. (1988): *Integrating Marker-Passing and Problem Solving: A Spreading Activation Approach to Improved Choice in Planning.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hillis, W. D. (1985). *The Connection Machine.* Cambridge, MA: The MIT Press.

Kitano, H., Tomabechi, H., & Levin, L. (1989). Ambiguity resolution in DMTRANS Plus. *Proceedings of the Fourth Conference of the European Chapter of the Association of Computational Linguistics.* Manchester Univ. Press.

Lange, T. (1990). Simulation of heterogeneous neural networks on serial and parallel machines. *Parallel Computing, 14,* 287-303.

Lange, T. (1992): Lexical and Pragmatic Disambiguation and Reinterpretation in Connectionist Networks. *International Journal of Man-Machine Studies, 36,* 191-220.

Lange, T. & Dyer, M. G. (1989): High-Level Inferencing in a Connectionist Network. *Connection Science, 1 (2),* 181-217.

Lange, T., Hodges, J. B., Fuenmayor, M., & Belyaev, L. (1989). DESCARTES: Development environment for simulating hybrid connectionist architectures. In *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society,* p. 698-705. Hillsdale, NJ: Lawrence Erlbaum.

Lange, T. & Wharton, C. M. (in press). REMIND: Retrieval from episodic memory by inferencing and disambiguation. To appear in J. Barnden and K. Holyoak (eds.), *Advances in connectionist and neural computation theory, volume II: Analogical connections.* Norwood, NJ: Ablex.

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

Norvig, P. (1989): Marker Passing as a Weak Method for Text Inferencing. *Cognitive Science, 13 (4),* p. 569-620.

Riesbeck, C. K., & Schank, R. C. (1989). *Inside case-based reasoning.* Hillsdale, NJ: Lawrence Erlbaum.

Shastri, L., & Ajjanagadde, V. (in press). From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. To appear in *Behavioral and Brain Sciences.*

St. John, M. F. (1990): The Story Gestalt: A model of knowledge intensive processes in text comprehension. *Cognitive Science, 16,* 271-306.

Sun, R. (in press). Beyond associative memories: Logics and variables in connectionist models. To appear in *Information Sciences.*

Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence,* 46, 259-310.

Touretzky, D. S. (1990): Connectionism and Compositional Semantics. In J. A. Barnden and J. B. Pollack (eds.), *Advances in Connectionist and Neural Computation Theory.* Norwood, NJ: Ablex.

Touretzky, D. S. & Hinton, G. E. (1988): A Distributed Connectionist Production System. *Cognitive Science, 12 (3),* p. 423-466.

Waltz, D. & Pollack, J. (1985): Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation. *Cognitive Science, 9 (1),* p. 51-74.