PROCEEDINGS OF THE INTERNATIONAL CONFERENCE
ON FIFTH GENERATION COMPUTER SYSTEMS 1984,
edited by ICOT.

PARALLEL INTERPRETATION OF NATURAL LANGUAGE

Jordan B. Pollack
Coordinated Science Laboratory
University of Illinois at
Urbana-Champaign
1101 West Springfield Avenue
Urbana, IL 61801 U.S.A.

David L. Waltz
Thinking Machines Corporation
245 First Street
Cambridge, MA 02142 U.S.A.
and Brandeis University
Waltham, MA 02254 U.S.A.

ABSTRACT

This is a description of research in developing a natural language processing system with modular knowledge sources but strongly interactive processing. The system offers insights into a variety of linguistic phenomena and allows easy testing of a variety of hypotheses. Language interpretation takes place on a activation network which is dynamically created from input, recent context, and long-term knowledge. Initially ambiguous and unstable, the network settles on a single interpretation, using a parallel, analog relaxation process. We also describe a parallel model for the representation of context and of the priming of concepts. Examples illustrating contextual influence on meaning interpretation and "semantic garden path" sentence processing are included. Furthermore, our model has been designed with the constraints of New Generation Computing in mind, and we describe our first pass at the architectural design of a message-merging communications network which implements the relaxation process.

1 INTRODUCTION

The interpretation of natural language requires the cooperative application of many kinds of knowledge, both language specific knowledge about word use, word order and phrase structure, and "real-world" knowledge about stereotypical situations, events, roles, contexts, and so on. And even though these knowledge systems are nearly decomposable, enabling the circumscription of individual knowledge areas for scrutiny, this decomposability does not easily extend into the realm of computation; that is, one cannot construct a psychologically realistic natural language processor by merely conjoining various knowledge-specific processing modules serially or hierarchically.

We offer instead a model based on the integration of independent syntactic, semantic, lexical, contextual, and pragmatic knowledge sources via spreading activation and lateral inhibition links. Figure 1 shows part of the network that is activated when the sentence

(S1) John shot some bucks.

is encountered. Links with arrows are activating, while those with circles are inhibiting. Mutual inhibition links between two nodes allow only one of the nodes to remain active for any duration. (However, both nodes may be simultaneously inactive.) Mutual inhibition links are generally placed between nodes that represent mutually incompatible interpretations, while mutual activation links join compatible ones. If the context in which this sentence occurs has included reference to "gambling," only the shaded nodes of Figure 1(a) remain active after relaxation of the network. If, on the other hand, "hunting" has been primed, only the shaded nodes shown in Figure 1(b) will remain active.

Notice that the "decision" made by the system integrates syntactic, semantic, and contextual knowledge: the fact that "some bucks" is a legal noun phrase is a factor in killing the readings of "bucks" as a verb; the fact that "hunting" is associated with both the "fire" meaning of "shot" and the "deer" meaning of "bucks" leads to the activation of the coalition of nodes shown in Figure 1(b); and so on. At the same time, the knowledge is discrete, and easy to add or modify. In this model of processing, decisions are spread out over time, allowing various knowledge sources to be brought to bear on the elements of the interpretation process. This is a radical departure from natural language processing models based on the convenient decision procedures provided by conventional programming languages.

Our program operates by (1) constructing a graph with weighted nodes and links from a sentence, and (2) running an iterative operation which recomputes each node's activation level (i.e its weight) based on a function of its current value and the inner product of

This work is supported by the Office of Naval Research under contract N00014-75-C-0612.

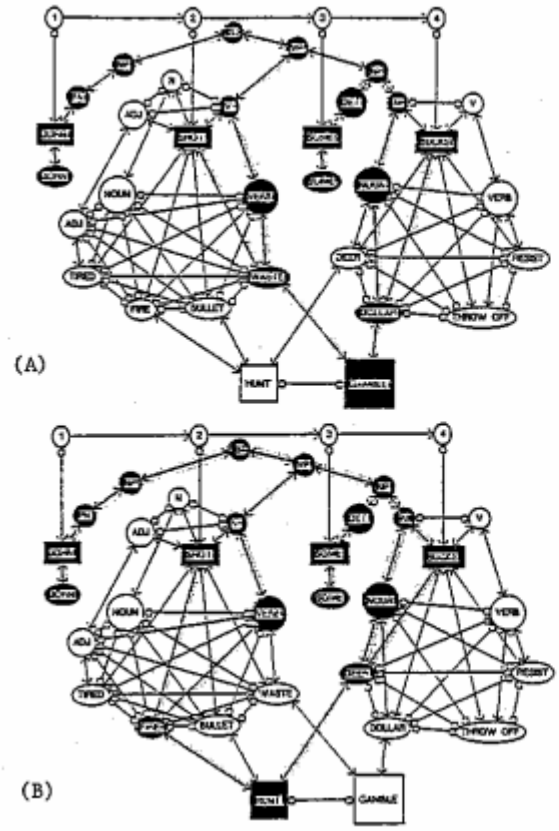


Figure 1: Two interpretations of "John shot some bucks."
(A) shows the result in the context of gambling, i.e. John wasted some money;
(B) shows the result in the context of hunting, i.e. John fired a gun at a deer.
Both examples required about 25 cycles to settle; in each case only a slight initial advantage was given to HUNT or GAMBLE. The numbered nodes control the arrival times of the words.

its links and the activation levels of its neighbors. For these examples, we are primarily interested in the behavior of the network, and not in the program that dynamically constructs the network. The syntactic portions of the networks in this paper were constructed by a chart parser (Kay 1973), while the semantic and contextual portions are permanently resident in memory. Initially all nodes are given zero weight, except for the nodes used to model context (e.g. "hunting" and "gambling"). Each activation link has a weight of .2 and each inhibition link has a weight of -.45. The iterative operation uses a proportional function to compute new weighting for nodes, similar to the one used by McClelland and Rumelhart (1980) in their interactive activation model.

The net effect of the program is that, over several iterations, a coalition of well-connected nodes will dominate, while the less fortunate nodes (those which are negatively connected to winners) will be suppressed. We exploit this behavior several ways in our system: by putting inhibitory links between nodes which represent well-formed phrases with shared constituents (which are, thus, mutually exclusive), we ensure that only one will survive. Similarly, there are inhibitory links between nodes representing different lexical categories (i.e. noun or verb) for the same word; between concept nodes representing different senses of the same word (i.e. submarine as a boat or as a sandwich); and between nodes representing conflicting case role interpretations. There are activation links between phrases and their constituents, between words and their different meanings, between roles and their fillers, and between corresponding syntactic and semantic interpretations.

2 MODELING PHENOMENOLOGY

Because our system operates in time, we are able to model effects that depend on context, and effects that depend on the arrival times of words. Consider the network shown in Figure 2, which shows three snapshots taken during the processing of the sentence[1] (due to Charniak (1983)):

(S2) The astronomer married a star.

Figure 2 includes three possible meanings for "star", namely (1) the featured player in dramatic acting, or (2) a celestial body, or 3) a pentagram. We presume that "astronomer" primes STAR by the path of strong links: astronomer -> ASTRONOMER -> ASTRONOMY -> CELESTIAL-BODY, but that MOVIE-STAR would be primed very little, if at all, because any activation of HUMAN via "astronomer" and "married" is spread fairly evenly among a vast number of other concepts (PHYSICIAN, PROFESSOR, etc.). When the word "star" is encountered, the meaning CELESTIAL-BODY is initially highly preferred, but eventually, since CELESTIAL-BODY is inanimate, whereas the object of MARRY should be human and animate, the MOVIE-STAR meaning of "star" wins out.

In Figure 2(d) we show the activation levels for CELESTIAL-BODY and MOVIE-STAR as functions of time. One can see that the activation of CELESTIAL-BODY is initially very high, and that only later does MOVIE-

[1] Our current system simulates the sequenced arrival of words by first constructing a complete network as shown in Figure 2 (using a chart parser), and then activating the node marked "1" at the upper left of Figures 2(A), (B), and (C). Node "1" activates both node "2" and the node "the". The node "the" deactivates node "1", and starts activation of the syntactic and semantic sections of the networks. Node "2" activates node "3" and the node "astronomer", etc. Eventually all words are activated, left-to-right. This simple sequencing mechanism is, thus, a stand-in for a more complete model of visual or auditory perception.

STAR catch up to and eventually dominate it. We argue that, if activation level is taken as a prime determinant of the contents of consciousness, then this model captures a common experience of people when hearing this sentence. This phenomenon is often reported as being humorous, and could be considered a kind of "semantic garden path". It should be emphasized that this behavior falls out of this model, and is not the result of juggling the weights until it works. In fact, the examples shown in this paper work in an essentially similar way over a broad range of link weightings.

3 CONTEXT: INTRODUCTION

Earlier (Figure 1) we used "context-setting" nodes such as "hunting" and "gambling" to prime particular word and phrase senses, in order to force appropriate interpretations of a noun phrase. There are, however, major problems that preclude the use of such context setting nodes as a solution to the problem of context-directed interpretation of language. A particular context-setting word, e.g. "hunting", may never have been explicitly mentioned earlier in a text or discourse, but may nonetheless be easily inferred by a reader or hearer. For example, preceding (S1) with:

(S3) John spent his weekend in the woods.

should suffice to induce the "hunting" context. Mention of such words or items as "outdoors", "hike", "campfire", "duck blind", "marksman", etc. ought to also prime a hearer appropriately, even though some of these words (e.g. "outdoors" and "hike") are more closely related to many other concepts than to "hunting." We are thus apparently faced with either (a) the need to infer the special context-setting concept "hunting", given any of the words or items above; or (b) the need to provide connections between each of the words or items and all the various word senses they prime. There is, however, a better alternative.

We propose that each concept should be represented not merely as a unitary node, but should in addition be associated with a set of "microfeatures" that serve both (a) to define the concepts, at least partially, and (b) to associate each concept with others that share its microfeatures. We propose a large set of microfeatures (on the order of thousands), each of which is potentially connected to every concept node in the system (potentially on the order of hundreds of thousands). Each concept is in fact connected to only some subset of the total set,

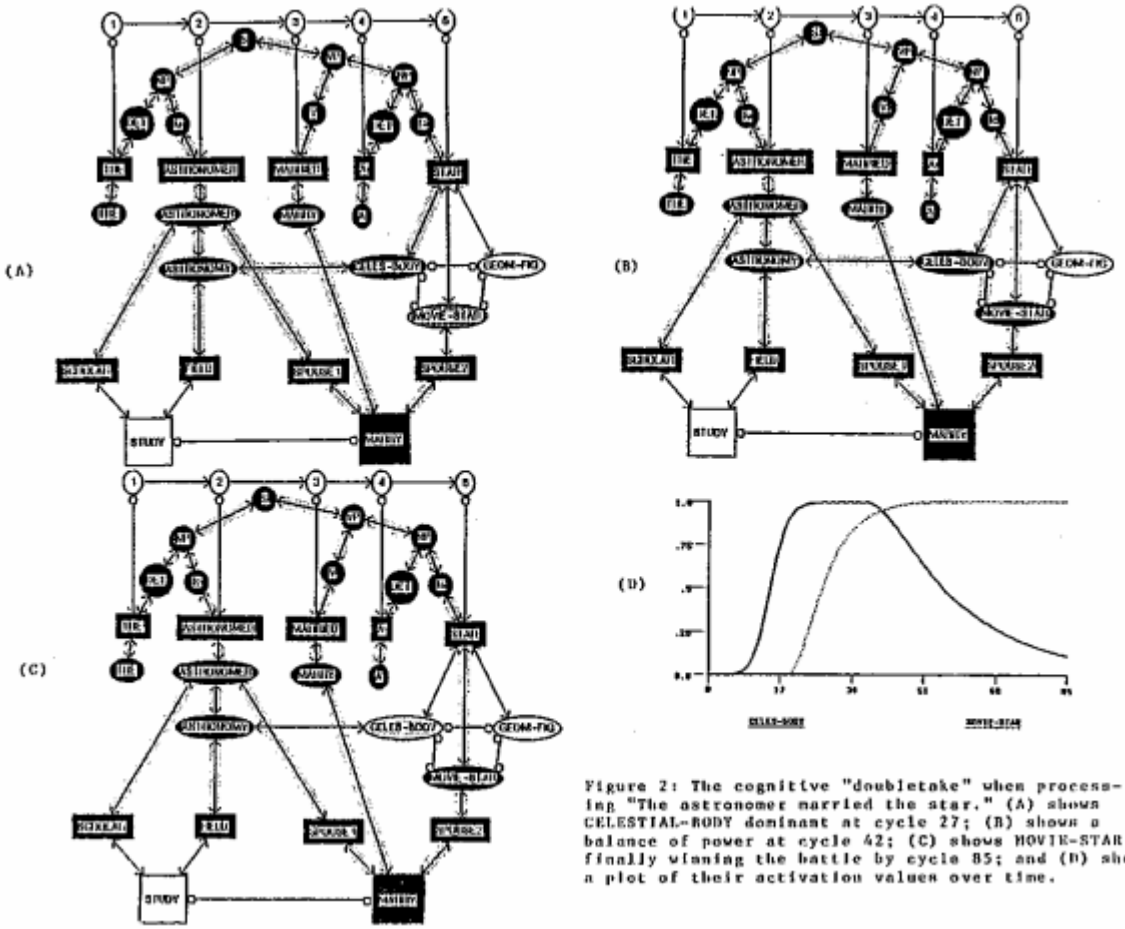


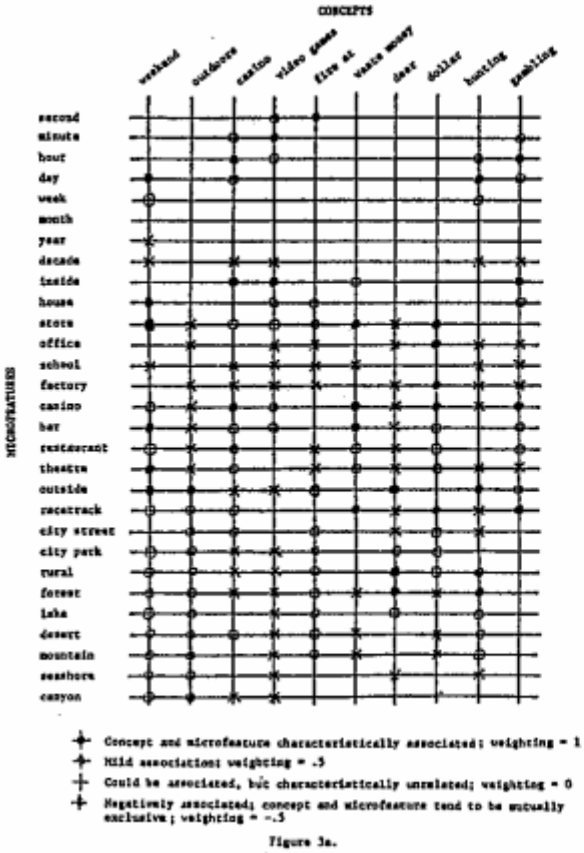
Figure 2: The cognitive "doubletake" when processing "The astronomer married the star." (A) shows CELESTIAL-BODY dominant at cycle 27; (B) shows a balance of power at cycle 42; (C) shows MOVIE-STAR finally winning the battle by cycle 85; and (D) shows a plot of their activation values over time.

via either bidirectional activation or bidirectional inhibition links. Closely related concepts have many microfeatures in common.

We suggest that microfeatures should be chosen on the basis of first principles to correspond to the major distinctions humans make about situations in the world, that is, distinctions we must make to survive and thrive. For example, some important microfeatures correspond to distinctions such as threatening/safe, animate/inanimate, edible/inedible, indoors/outdoors, good outcome/neutral outcome/bad outcome, moving/still, intentional/unintentional, or characteristic lengths of events (e.g., whether events require milliseconds, hours, or years). As in Hinton's (1981) model, hierarchies arise naturally, based on subsets of shared microfeatures, but are not the fundamental basis for organizing concepts in a semantic network, as in most AI models.

3.1 Microfeatures as a Priming Context - An Example

Let us see how microfeatures could help solve the problems presented by the example in Figure 1. Figure 3 shows a partial set of microfeatures, corresponding to temporal event length or location (setting) running horizontally. A small set of concepts relevant to our example is listed across the top. Solid circles denote strong connection of concepts to microfeatures, open circles, a weak connection, and crosses, a negative connection.



✚ Concept and microfeature characteristically associated; weighting = 1

✚ Mild association; weighting = .5

✚ Could be associated, but characteristically unrelated; weighting = 0

✚ Negatively associated; concept and microfeature tend to be mutually exclusive; weighting = -.5

Figure 3a.

A simple scoring scheme allows "weekend" and "outdoors" to appropriately prime concepts related to "fire at" and "deer" relative to "waste money" and "dollar," as well as the ability of "casino" or "video game" to induce an opposite priming effect, as shown in Figure 3(b). It is interesting to compare these effects with the effects of priming with "hunting" or "gambling" directly. No relaxation was used, though it obviously could be (i.e. a concept could activate microfeatures, priming other concepts, and then the primed concepts could change the activation of the microfeatures, in turn activating new concepts and eventually settling down.[2] We have been experimenting with a number of possible weighting and propagation schemes, and have built up a much larger matrix than the one shown in Figure 3.

4 RELATED WORK

There are many research projects which are very much in the same spirit as ours in addition to ones mentioned already in this paper. Beginning in the early 1970's, Schank argued that semantics, not syntax, should have the central role in both theories and programs for natural language processing; Riesbeck's parser for MARGIE (Schank et al. 1973) has a clear relationship to the model proposed here. Steven Small (1980) was another worker in AI to question the traditional serial integration of language pro-

PRIMING CONCEPTS	PRIMED CONCEPTS			
	Fire-at	Waste	Deer	Dollar
Weekend	.41	.55	0	.46
Outdoors	.41	0	.44	.08
Casino	.05	.59	0	.42
Video Games	.18	.36	0	.19
Weekend + Outdoors	.41	.07	.25	.12
Hunting	.36	0	.50	0
Gambling	.09	.59	0	.38

Fraction of Maximum Possible Score

Figure 3b: Instantaneous priming effects on concepts; microfeatures start at 0, and undergo a single priming cycle.

Figure 3: This figure illustrates the use of microfeatures to provide contextual priming. At any given time, microfeatures will display some pattern of activation. Each concept has an induced activation level as a result of the microfeature activation values. The microfeature activations are modified whenever a concept is primed.

For our example, assume "weekend" is primed, with all microfeatures initially at 0. The top line of Figure 3b shows the activation levels of concepts, where the number represents a fraction of the maximum possible activation for that concept. These values prime various word sense nodes differentially.

cessing. He suggested that rather than having separate modules for syntax and semantics, each word was an expert in interpreting its own meaning and role in context. Following on that work, Gary Cottrell is re-casting word-sense selection into a connectionist framework, and his work is very closely related to our own (Cottrell et al. 1983). Mark Jones (1983) is also working on parsing with spreading activation, but of the digital kind.

Other work has set integrated parsing into the production system framework. BORIS (Dyer 1982) uses a lexically-based demon-driven production system to read stories and answer questions about them. The READER system (Thibadeau et al. 1982) is a multi-level parallel production system which models chronometric data, i.e. data on how long humans visually fixate on each word while reading.

Another interesting approach to language integration is taken by Hendler and Phillips (1981), who are using a message-passing ACTOR (Hewitt 1976) system to model the interactions between syntax, semantics and pragmatics. Other work that has influenced our research includes the spreading activation work by Ortony and Radin (1983), based on a network of free associations to English words.

5 ARCHITECTURAL CONSIDERATIONS

Our work, and, in general, other work in connectionist modelling (Feldman and Ballard 1982; Rumelhart and McClelland 1980) is particularly well-suited for implementation on New Generation parallel computers. Unlike cognitive models based on parallel production systems such as HEARSAY II (Lesser et al. 1975) or READER (Thibadeau et al. 1982), in which concurrent access to the "blackboard" is a bottleneck permitting only small speed-ups (Fennel and Lesser 1977), connectionist models permit a speedup proportional to the number of processors.

There are both advantages and disadvantages of these models with respect to the communications costs in a parallel system. One disadvantage is that since a basic processing cycle consists of a whole barrage of messages crossing the network, message-passing architectures with indeterminate delays are inappropriate. One advantage is that since each message is a quantitative value which is ultimately to be summed, we can distribute the addition through the network. We have designed two such communication networks for modelling activation networks in parallel using the concept of message merging processors. In the first design (Pollack 1982), each activation node corresponds to a NMOS cell, which contains memory for its activation level, an ALU and special-purpose sorting shift-registers for its links. The cells are laid out in the simplest geometry -- a linear array, and processing takes place in three stages: First, the activation and inhibition links, which are composed of a relative destination and magnitude, are multiplied by the current activation level and loaded into shift registers; second, the full barrage of messages is forwarded through the network in a constant number of very small shifting cycles[3]; and third, the activation levels are recomputed. The second design (Debrunner 1982) generalized this process into a two-dimensional topology.

6 CONCLUSION

We have not actually built the hardware, but continue to refine the natural language model, always keeping the constraints of VLSI implementation in mind. We have been developing our programs in LISP,[4] but we intend to implement them on Connection Machine™ hardware (Hillis 1981) when it becomes available.

Using spreading activation and lateral inhibition enables a good framework for embedding comprehension phenomena which cannot even be approached with binary serial models. While we have not discussed them here, we have explored ties to psychological and linguistic results and theories; these are reported in Waltz and Pollack (1984). There we show that structural preferences

---

[2] We have tried hard to be fair in constructing Figure 3(a), for example priming with "outdoor" rather than "woods," and including links between "casino" and "desert" to acknowledge Las Vegas. Time periods characterize event lengths. Locations are to be taken as settings or surrounds, not objects. All links are clearly culturally dependent though, we think, roughly in accord with current middle-class American language usage.

[3] The shift registers keep the messages sorted to send out the longest one first, and combine messages with the same destination. The result is that the length of the longest message decreases by 1 every shift cycle, leading to a constant time (shift-time * length(longest message)).

[4] Our system is currently implemented on both the Xerox 1108 (dandelion) and the Symbolics 3600. The Symbolics version interfaces with a complete dictionary (Webster's Seventh New Collegiate), while the Xerox version has a small, hand built lexicon. The grammars for both systems are still quite incomplete. Semantic portions of both systems are hand built, and also rather small (< 100 words). This work should be viewed as exploratory and suggestive; our system is not yet a practical alternative for real applications.

™ A product of Thinking Machines Corporation, Cambridge, MA, USA.

such as Minimal Attachment (Frazier 1979) can be understood as side-effects of, rather than as strategies for, a syntactic processor; current hypotheses about lexical disambiguation in context (Swinney 1979; Seidenberg et al. 1980) can nicely fit into a model with lateral inhibition; it could not be accounted for by activation alone. Garden-paths at different levels of processing can be explained by the breakdown of a common approximate consistent labeling algorithm -- Lateral Inhibition -- the "Universal Will to Disambiguate."

REFERENCES

Charniak, E. Passing Markers: A Theory of Contextual Influence in Language Comprehension. Cognitive Science 7, 3, 171-190 1983.

Cottrell, G.W. and Small, S.L. A Connectionist Scheme for Modelling Word Sense Disambiguation. Cognition and Brain Theory 6, 1, 89-120 1983.

Debrunner, C. A Two-Dimensional Activation Cell. Working Paper 46, Coordinated Science Laboratory, University of Illinois, Urbana, December 1982.

Dyer, M. In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension. Yale Computer Science Research Report 219, May 1982.

Feldman J.A. and Ballard, D.H. Connectionist Models and Their Properties. Cognitive Science 6, 3, 205-254 1982.

Fennel, R.D. and Lesser, V.R. Parallelism in AI Problem-Solving: A Case Study of HEARSAY II. IEEE Transactions on Computers, 98-111, February 1977.

Frazier, L. On Comprehending Sentences: Syntactic Parsing Strategies. Indiana University Linguistics Club 1979.

Hendler, J. and Phillips, B. A Flexible Control Structure for the Conceptual Analysis of Natural Language Using Message Passing. TR-08-81-03, Texas Instruments, Dallas 1981.

Hewitt, C. Viewing Control Structures as Patterns of Passing Messages. AI Memo 410, MIT AI Lab 1976.

Hillis, W.D. The Connection Machine (Computer Architecture for the New Wave), AI Memo 646, MIT AI Lab 1981.

Hinton G.E. Implementing Semantic Networks in Parallel Hardware. Parallel Models of Associative Memory, G.E. Hinton and J.A. Anderson (eds.). Lawrence Erlbaum Associates, Hillsdale 1981.

Jones, M.A. Activation Based Parsing. Proc. IJCAI, Karlsruhe, West Germany, 678-682 1983.

Kay, M. The MIND System. Natural Language Processing, Rustin (ed.). Algorithmics Press, New York 1973.

McClelland, J.L. and Rumelhart, D.E., An Interactive Activation Model of the Effect of Context in Perception. TR 91, Center for Human Information Processing, UCSD 1980.

Pollack, J.B. An Activation/Inhibition Network Cell. Working Paper 31, Coordinated Science Laboratory, University of Illinois, Urbana 1982.

Ortony, A. and Radin, D. SAPIENS: Spreading Activation Processor for Information Encoded in Network Structures. Tech. Rept. 296, Center for the Study of Reading, University of Illinois, Urbana, October 1983.

Riesbeck, C. and Schank, R.C. Comprehension by Computer: Expectation-Based Analysis of Sentences in Context. Research Report 78, Computer Science Dept., Yale University 1976.

Schank, R.C., Goldman, N., Rieger, C. and Riesbeck, C. MARGIE: Memory, Analysis, Response Generation and Inference in English. Proc. IJCAI, Stanford University, 255-262, 1973.

Seidenberg, M.S., Tanenhaus, M.K. and Leiman, J.M. The Time Course of Lexical Ambiguity Resolution in Context. TR 164, Center for the Study of Reading, University of Illinois, Urbana, March 1980.

Small, S. Word Expert Parsing: A Theory of Distributed Word-Based Natural Language Understanding. TR-954, Computer Science Dept., University of Maryland 1980.

Swinney, D.A. Lexical Access During Sentence Comprehension: (Re)consideration of Context Effects. Journal of Verbal Learning and Verbal Behavior 18, 645-659 1979.

Thibadeau, R., Just, M.A. and Carpenter, P.A. A Model of the Time Course and Content of Reading. Cognitive Science 6, 2, 157-203 1982.

Waltz, D.L. and Pollack, J.B. Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation. Cognitive Science 1984, to appear.