COGNITIVE SCIENCE 9, 51-74 (1985)

Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation*

DAVID L. WALTZ
JORDAN B. POLLACK
Coordinated Science Laboratory
University of Illinois

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, among other issues, are included.

INTRODUCTION

The Problem of Integration

The interpretation of natural language requires the cooperative application of many systems of knowledge, both language specific knowledge about word use, word order and phrase structure, and "real-world" knowledge

* This work is supported by the Office of Naval Research under Contract N00014-75-C-0612. This article has benefited from discussion with various people at the University of Illinois, including Bill Brewer, Gerald DeJong, Rick Dinitz, Marcy Dorfman, David Farwell, Georgia Green, La Raw Maran, Jerry Morgan, Andrew Ortony, and, especially, The Great Lorenzo. Thanks also to Cathy Cassells for limitless editing and typing, and to Pollack's program for printing the pictures.

Correspondence and requests for reprints can be sent to David L. Waltz at: Thinking Machines Corporation, 245 First St., Cambridge, MA 02142, or Computer Science Department, Brandeis University, Waltham, MA 02254.

about stereo-typical 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.

These particular forms of process integration, which happen to be convenient to use on modern computers, turn out to have a profound effect on the mind of the modeler, as Pylyshyn (1980, p. 124) points out:

> Now, what is typically overlooked when we [use a computational system as a cognitive model] is the extent to which the class of algorithms that can even be considered is conditioned by the assumptions we make regarding what basic operations are possible, how these may interact, how operations are sequenced, what data structures are possible, and so on. Such assumptions are an intrinsic part of our choice of descriptive formalism.

Ambiguity. Convenient processing assumptions lead to problems in building models for cognition. Consider ambiguity, perhaps the most ubiquitous problem in natural language processing. Humans experience an increased processing load with ambiguous language (Mackay, 1966), which suggests that humans compute multiple readings (at least in some sense). However, the "serial frame of mind" allows basically two approaches for dealing with ambiguous sentences: *backtracking* as used in Augmented Transition Networks (Woods, 1970), or *delay*, as used in Marcus' (1980) "wait and see" parser. And although lexical access appears to be an automatic process cotemporal with syntactic and semantic processing (Marslen-Wilson & Tyler, 1980) most natural language systems still work with small dictionaries and simple ad-hoc heuristics which choose word meanings before assigning structure.[1]

Single Interpretation. Another interesting phenomenon in language interpretation is that humans can usually entertain only one interpretation of an ambiguous sentence at a time, but can easily "flip" between interpretations. Consider the following short sentences which can be interpreted either as statements or commands:

(S1) Trust shrinks.
(S2) Respect remains.
(S3) Exercise smarts.

The fact that we can interpret the first sentence either as a general statement about dwindling confidence or as advice to place one's faith in psychiatrists,

[1] This, finally, is changing. See Small (1980) or Charniak (1983).

suggests that the human capacity to disambiguate language is not unlike the faculties involved in visual disambiguation—like foreground/background perception of the Necker Cube.

Comprehension Errors. Errors in comprehension, like Garden Path Sentences have in the past been explained by purely structural principles like early closure (Kimball, 1973) or Minimal Attachment and Right Association (Frazier, 1979), or by the breakdown of simple and limited serial mechanisms (Marcus, 1980; Milne, 1982; Shieber, 1983). However, they have more complete and natural explanations as side-effects of strongly interacting processes as we demonstrate in the section on Errors in Comprehension, p. 61. Garden path effects can occur at all levels of language processing. Consider the following sentences:

(S4) The astronomer married the star. (Charniak, 1983)
(S5) The plumber filled his pipe.
(S6) The sailor ate a submarine.

Readers usually report these as "temporarily anomalous," i.e., as a "semantic garden path sentence." A plausible explanation for the cognitive doubletake (and mild humor) caused by the first sentence is that the priming power of "astronomer" on the wrong sense of "star" is initially stronger than the logical power of case frame selectional restrictions; in the end, the selectional restrictions force the interpretation of "star" as a person.

Nongrammatical Text. People are able to interpret nongrammatical language, whether it is naturally occurring (due to poor grammar, foreign speakers, noise interference, interruptions, self-corrections, etc.) Most work on this topic has taken the approach of relaxing certain contraints in the parsing process—in the LSP project (Sager, 1981), a failed parse was retried without agreement constraints on syntactic features; in the PLANES project (Waltz, 1978), a semantic grammar was used which accepted very ungrammatical input as meaningful. Others have formalized the notion of constraint relaxation for handling ill-formed input (Goodman, 1984; Kwasny & Sondheimer, 1981). We believe that this ability in humans, to semi-independently judge meaningfulness and grammaticality, is yet more evidence of the modularity of knowledge but the integration of processing.

Integrating Knowledge Sources

All these phenomena indicate the need for a theory of language processing which posits, instead of the simple passing of semi-complete results between processing components, strong interaction between those components; so strong, in fact, that all decisions are interdependent. We believe that most theories of language processing advanced over the years have been seriously

flawed because they have drawn on a limited set of computational ideas which cannot effectively deal with interdependent decisions, and because of peculiarities of English and the history of linguistics research which have led to the assumption of the autonomy of syntax in natural language understanding.

These observations are, of course, not entirely new. In the early 1970s, Schank argued that semantics, not syntax should have the central role in programmed theories of natural language processing (Schank, Goldman, Rieger, & Riesbeck, 1973; Riesbeck & Schank, 1976). Steven Small (1980) was another worker in AI who questioned the traditional serial integration of language processing. Small suggested that rather than having separate modules for syntax and semantics, each word should be its own expert, and built a system of interacting discrimination nets, reminiscent of Hewitt's work on actor formalisms (Hewitt, 1976). Cottrell and Small (1983) have published research on interacting distributed processing of word senses, case roles, and semantic markers, very similar in spirit to our effort (see section titled "Case Frames and Selectional Restrictions," pp. 64–65). And, of course, the HEARSAY II speech understanding system (Fennel & Lesser, 1977) used a parallel production system for integrating multiple knowledge sources.

STRUCTURES OF THE MODEL

We wish to go further than these models of integrated processing; we want a machine to make its interdependent decisions smoothly—we want a system which runs like an ecological system, rather than like a Rube Goldberg device. We have chosen to work with the twin processes of *Spreading Activation* and *Lateral Inhibition* in which decisions are spread out over time, allowing various knowledge sources to be brought to bear on elements of the interpretation.

Spreading Activation and Lateral Inhibition

The term "Spreading Activation" has been used to describe many different programs and models, but all can be basically divided into two classes. *Digital Spreading Activation* is a class of marker-passing algorithms which perform a breadth-first search for shortest paths on a relational network. *Analog Spreading Activation* takes place on a weighted network of associations, where "activation energy" is distributed over the network based on some mathematical function of the strength of connections. As an example of the digital kind, Quillian (1968) describes a technique for finding relationships between two concepts stored in a semantic network by repeatedly marking

adjacent nodes with "activation tags" containing a path back to the source concept. As examples of the analog kind, Collins and Loftus (1975) model semantic priming on a network with decaying activation, and, more recently, McClelland and Rumelhart (1980) model word recognition on an analog spreading activation network.

Both forms of spreading activation suffer from the danger of "overkill." In the digital form, this means the problem of false positives, where a very large number of uninteresting paths may be found.[2] The analog form of spreading activation has the potential problem of "heat death," where the entire network becomes uniformly activated. This can be handled with some form of damping or decay, or via lateral inhibition, which spreads negative energy just as spreading activation spreads positive energy. With damping or decay, weights must be carefully chosen to avoid having too many or too few portions of the network active. Lateral inhibition seems to have fewer disadvantages, and is our method of choice.

Besides effectively dealing with the problem of overkill, lateral inhibition is useful for the coordination of distributed decisions. According to Feldman:

> Lateral inhibition at lower organizational levels is one of the most ubiquitous information-processing mechanisms in animals: it is essential that opposing action systems do not execute simultaneously. Low-level visual processing makes very heavy use of mutual lateral inhibition, and this appears to be true for other sensory systems as well (1981, p. 52).

Consider a graph with weighted nodes and links, and 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 its links the activation levels of its neighbors. An activation (positive) link between a pair of nodes will cause them to support each other while an inhibition (negative) link will attempt to allow only one of the pair to remain active at any given time.[3] The net effect 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 parser: 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

[2] Quillian recognized that this was a problem for preposition words in his network, and used a simple heuristic of avoiding paths through dense clusters to circumvent it (1968, p. 156). Charniak (1983, p. 188) discusses a similar heuristic.

[3] The pair may, under certain circumstances, balance each other; both may have zero activation simultaneously.

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. The net effect is that, over several iterations, a coalition of nodes representing a consistent interpretation will dominate, while the less fortunate nodes will be suppressed.

Integration by Parts

While we have not yet built an entire language processing system, we have done several computational "experiments," modeling various components of comprehension, as well as certain interactions between them. We will describe the different components we have modeled with activation and inhibition and how they fit together. The figures in this section have nodes represented by named and shaped shapes—the darker the shade, the higher the activation level. Activation links are shown with arrowheads, while inhibition links end with small circles. All the figures are snapshots taken of a system which runs a proportional update activation function, the same as McClelland and Rumelhart's (1980) scheme, on a range from 0.0 to 1.0 without decay. All activation links are at $+.2$ and all inhibition links at $-.45$.

Lexical Access and Priming. One of the earliest uses for spreading activation was to model semantic priming (Collins & Loftus, 1975; Ortony & Radin, 1983). It is very natural to think in terms of weighted connections between related concepts as being responsible for the "associations" that drive the facilitation of word meanings in context. AI models, with the exception of Small's (1980) Word Expert Parser, have basically ignored this aspect of language processing as "lower than syntax," and have usually called a subroutine to pick out the lexical category and meaning of a word upon demand.

Both approaches are a little off. Recent evidence about lexical access (Seidenberg, Tanenhaus, & Leiman, 1980; Swinney, 1979) shows that when a word is encountered, all its meanings are facilitated (phase 1) but then rapidly, the local semantic and syntactic context eliminate all but the "correct" one (phase 2). Spreading activation can easily be responsible for phase 1, but it provides no way to achieve phase 2; opportunistic selection by subroutine call can, in many cases, pick out the right lexical item, but does not accurately model the automatic process humans do. Using lateral inhibition between competing word senses, and between word senses and the local syntactic and semantic context, can effectively simulate the two stage hypothesis without two processing stages.

Figure 1 shows a four-level activation network for the sentence:

(S7) John shot some bucks.

Figure 1a shows the initial state, and 1b shows the network after about 50 cycles. The first level shows a syntactic parse tree for the sentence, the second level shows the input words, the third level shows a cluster of meanings for each word with mutual inhibition links between all the meanings, and crossed activation and inhibition links between the lexical categories and the meanings. For example, the input word "shot" is connected to four senses:

1. TIRED—an adjective, approximately meaning[4] "worn out"
2. FIRE—a past-tense verb, approximately meaning "to shoot with gun"
3. BULLET—a noun, approximately meaning "unit of ammunition"
4. WASTE—a verb, approximately meaning "squander a resource"

The fourth level, in large italics, represents an image of a context system which is discussed in the section titled "Context: Introduction" on p. 65. Its purpose is to coordinate sense selection along the sentence.

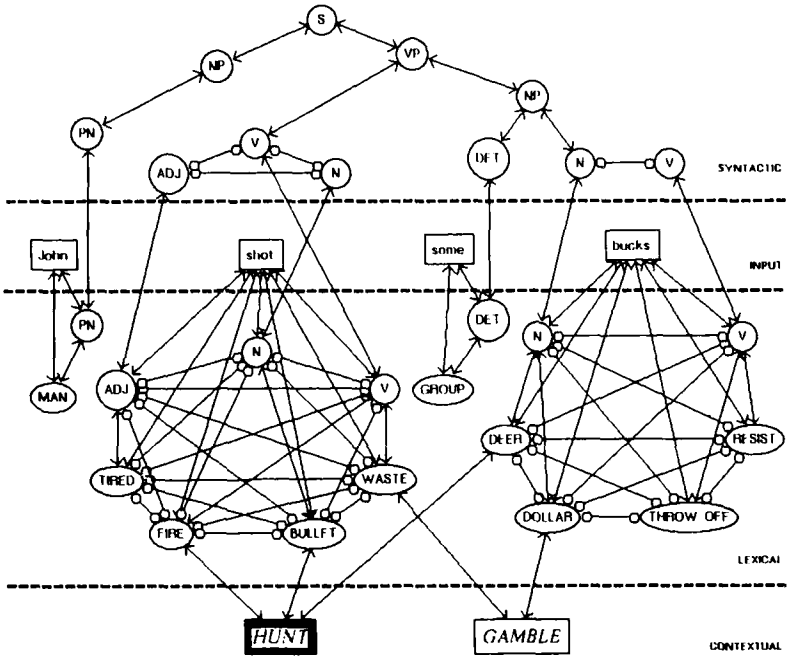

Figure 1a. Initial Network for "John shot some bucks" in the context of hunting.

[4] The nodes in this diagram, and in this paper in general are intended to have a schematic internal representation. We see the behavior of our networds as a coarse description of an even larger, more parallel representation of schemata and semantic features.

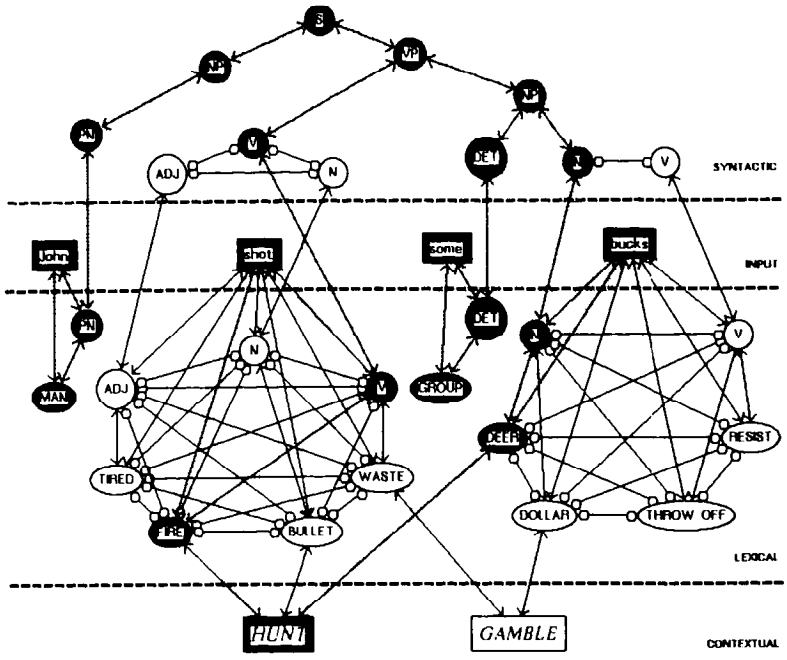

Figure 1b. Network after 50 cycles of spreading activation and lateral inhibition.

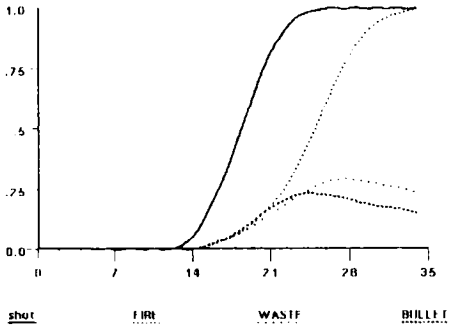

Figure 1c. Graph of the activation profile of "shot" and several of its senses. The horizontal axis represents time, in cycles (and no claims are being made at this point for a mapping to milliseconds), and the vertical axis represents activation levels.

The activation of this network (by an unshown auxiliary network which sequences the words) yields an interesting profile of the processing of "John shot some bucks" in the context of hunting. The graph in Figure 1b shows the activation profiles over time of the word "shot" and some of its meanings. Meanings are activated immediately after the word, and the syntactic demand for a verb, along with the contextual pressure of hunting, lead to the rapid demise of all but the "FIRE" sense of "shot."

Syntax—Autonomy and Integration. It is clear that humans can make relatively autonomous judgements about the grammaticality and meaningfulness of sentences. This is clearly demonstrated by the ability to assign structure to nonsense sentences such as:

(S8) Colorless green ideas sleep furiously.[5]

as well as to assign some meaning to ungrammatical strings. The autonomy of syntax is important because it allows a very complex element of natural language to be studied in isolation from other elements. Unfortunately, it has been misunderstood at an implementational level, with programmers having drawn the wrong conclusion: that natural language can be processed by a "syntactic faculty"[6] that assigns structure before meaning. This misunderstanding and the failures that accompanied it, along with the early successes of meaning-primitive based systems (Schank et al., 1973), have led many AI researchers to assume a "rejectionist" position with regard to syntax.[7]

Obviously, syntax is not the framework upon which an entire NLP system should be based, but neither should it be dispersed into the farthest reaches of subroutine calls. We put syntax on an equal footing with other sources of language knowledge. Notice that the misunderstanding of syntactic autonomy stems from the intellectual bottleneck of serial processing discussed earlier—when computations must be serialized, some decisions must be made before others. In a parallel and strongly interactive framework, syntax can be integrated in such a way as to allow relatively independent judgements of grammaticality, as well as to influence and be influenced by judgements of meaningfulness.

Our approach to relatively independent syntactic processor (Pollack & Waltz, 1982) is based on merging ideas from breadth-first chart parsing (Kay, 1973) with parallel relaxation by lateral inhibition. The output from a chart parser for a context-free grammar[8] is taken to be a network with activation links between mother and daughter nodes, and inhibition links between "in-laws"—nodes which both dominate a common daughter. When the network is iterated, it seeks an equilibrium state with an active coalition of phrase-markers representing a well-formed parse.

Figure 2 shows two stable states of a network representing the parses of the sentence

(S9) John ate up the street.

[5] From Chomsky, (1957, p. 15).

[6] See Fodor (1982) for a history of faculty theories of mental abilities.

[7] Though a comeback of sorts seems to be underway—see Berwick, (1983).

[8] Although common wisdom says that CFG's are inadequate to represent the syntax of natural languages, recent work on Generalized Phrase Structure Grammars (Gazdar & Pullum, 1982) holds the promise of overcoming the inadequacies.

The network settles on one of the states based on the initial node and link weights, and on any external activation or inhibition applied to the nodes, i.e., from the lexical or semantic level. The important thing to note is that there is no "homunculus" searching for a consistent tree, just a local competition for superiority. The weighting scheme where the words are activated sequentially from left to right prefers to settle on the shortest tree it can find, Figure 2a.

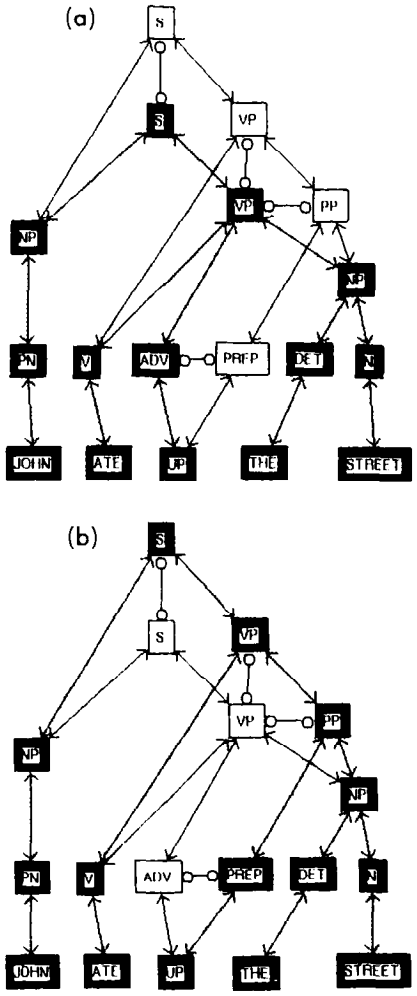

Figure 2. Stable coalitions of the ambiguous syntax of "John ate up the street." The reading of "up" in 2a (John ate all of the asphalt) is preferred but semantic information can easily overcome this preference.

There has been a great deal of research into exactly how the "human syntactic processor" works. Data on which decisions humans make in the structural interpretation of sentences in null contexts have generated many principles and strategies for parsing (Frazier, 1979; Ford, Bresnan, & Kaplan, 1982; Kimball, 1973), and this data has been used as validation of a whole host of grammatical formalisms and parsing mechanisms, including the Sausage Machine (Frazier & Fodor, 1978) ATN's (Wanner, 1980), and different varieties of deterministic parsers (Church, 1980; Marcus, 1980; Milne, 1982; Shieber, 1983). There are a number of problems with building a syntactic parser to explicitly encode these strategies as the starting point for a language understanding system. We list three major ones here. First, humans can understand sequences of words with ill-formed syntactic structure. Second, when beginning with a syntactic parser, there is no *good* way to smoothly integrate semantic and lexical strategies, so the different strategies are usually subjugated to the status of subroutines called by some control program of dubious psychological and linguistic credentials (Winograd, 1972). Third, since garden path sentences have been taken as validation of the structural principles, it would be good if they always caused humans to garden path. Not so, Crain and Steedman (1981) have shown how easy it is to prevent backtracking by adding context to garden path sentences. For example, if

(S10) Cotton is grown in several of the Southern States.

precedes

(S11) The cotton clothing is made of is grown in Mississippi.

the latter is no longer a garden path sentence.

Since structural preferences are apparently so much weaker than other contextual forces, is it not better to view them as side-effects of the organization of the syntactic processor, rather than as global guiding principles? In a parallel parser such as the one described above, preferences between competing phrases are a side-effect of lateral inhibition; the global guiding principle is the "universal will to disambiguate."

It is interesting to note that although our model is motivated by considerations of strong interaction, the syntactic modeling is closely related to the HOPE program of Gigley (1982) which was motivated by the neurolinguistic concerns of modeling aphasiac language degradation. Currently, however, we have no goals of lobotomizing our program to see how it degrades (but see Marcus, 1982 for one such experiment).

Errors in Comprehension. Because our system operates in time, we are able to model effects that depend on context, and effects that depend on the arrival time of words. Consider the network shown in Figure 3, which shows

three snapshots taken during the processing of a sentence that induces a "cognitive doubletake":

(S4) The astronomer married a star.

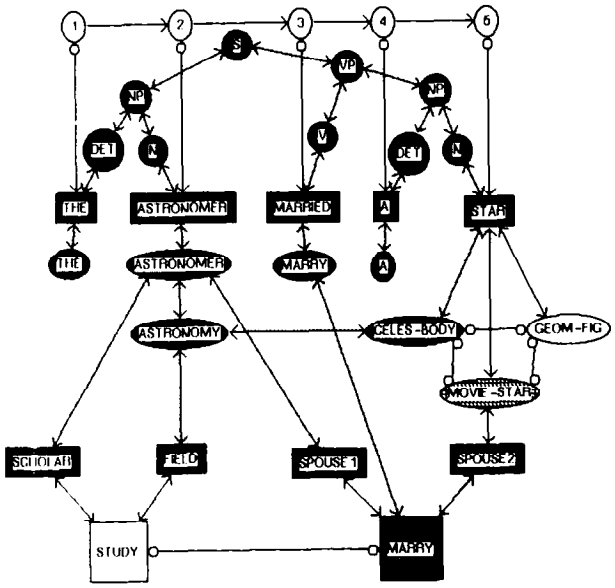

Figure 3a. Cycle 27 of "The astronomer married the star"; CELES-BODY seems to have won.

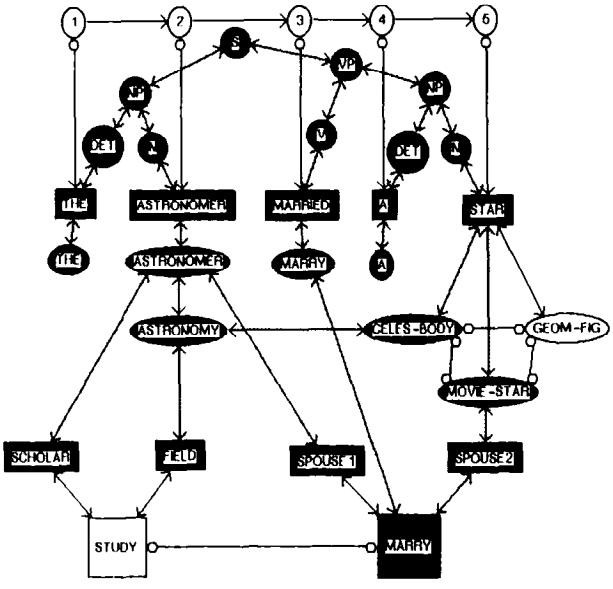

Figure 3b. Cycle 42; MOVIE-STAR has caught up.

Figure 3 includes three possible meanings for "star," namely (1) MOVIE-STAR—the featured player in dramatic acting, or (2) CELES-BODY—a celestial body, or 3) GEOM-FIG—a pentagram. We presume that "astronomer" primes CELES-BODY by the path of strong links: astronomer→ASTRONOMER→ASTRONOMY→CELES-BODY, but that MOVIE-STAR would be primed very little, if at all, because the its activation via the distributed context model would be very small. When the word "star" is encountered, the meaning CELES-BODY is initially highly preferred and seems to have won the competition (Figure 3a), but eventually, since CELES-

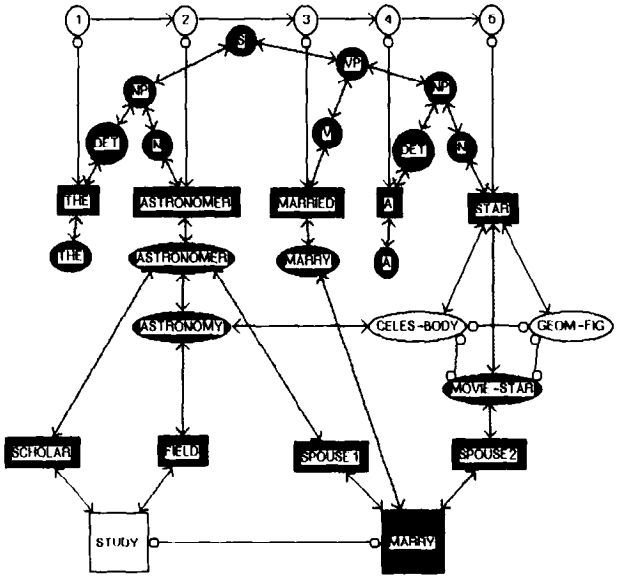

Figure 3c. Cycle 85; CELES-BODY has lost to MOVIE-STAR; consistency reigns.

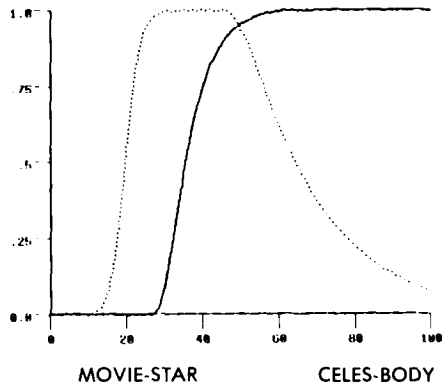

Figure 3d.

BODY is inanimate, whereas the object of MARRY should be human and animate, the MOVIE-STAR meaning of "star" catches up (Figure 3b) and wins out (Figure 3c).

In Figure 3d we show the activation levels for CELES-BODY and MOVIE-STAR as functions of time. One can see that the activation of CELES-BODY is initially very high, and that only later does MOVIE-STAR catch up to and eventually dominate it. We argue that, if activation level is taken as a prime determinant of the contexts 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.

Case Frames and Selectional Restrictions. Figure 3 includes some large square nodes at the bottom. These large boxes, representing case frames activated by "ASTRONOMER" and "MARRY", actually correspond to substantial structures of nodes and links, in basic agreement with the "exploded case frame" of Cottrell and Small (1983). Specifically, each case frame includes role slots, specific to the case frame: "MARRY" is attached via activation links to "MARRY-AGENT" and "MARRY-OBJECT" nodes, and these are each attached to semantic marker nodes for "HUMAN."

A scheme also is required to attach specific words, e.g., "astronomer," to roles as well as to the contexts of long-term memory. Finally, if the sentence's meaning is to be remembered, a scheme is necessary for dynamically connecting the active semantic and pragmatic nodes to long-term memory. All these schemes, in our view, require that there be (a) a way of collecting all active nodes, (b) a way of attaching the nodes to some other node (or nodes) unique to the set of active nodes, such that (c) the set of active nodes could be reactivated by activity of subsets of the set of nodes. Two methods for accomplishing this are Minsky's use of "K-lines" (Minsky, 1980) and Hinton's use of "microfeatures" (Hinton, 1981). The basic idea in both cases is that some node ("agent" for Minsky) or portion of a state vector (Hinton) is associated with specific activated nodes, and either bidirectional links (Minsky) or autoassociative hardware (Hinton) is used to recover the whole from any sufficiently large part.

Furthermore, we believe that each basic meaning node for a verb should itself be a composite structure: "eat," for instance, can be decomposed into schemas for moving food to a mouth, chewing, swallowing, and so on. Some ideas along these lines are explored in the section of DeJong and Waltz (1983) on "event shape diagrams." These diagrams can be viewed as plotting activation levels of such "microschemas." Ultimately, schemas

should connect to even more detailed mechanisms for producing the experience of mental images (Waltz, 1979).

Context: Introduction

Earlier (in Figure 1) we used "context-setting" nodes such as "HUNT" and "GAMBLE" 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 the text or discourse, but may nonetheless be easily inferred by a reader or hearer. For example, preceding (S1) with:

(S12) 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," and so on, 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 the concept with others that share its microfeatures. We propose a large set of microfeatures (on the order of a thousand), 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, via either bidirectional activation or bidirectional inhibition links. Closely related concepts have many microfeatures in common. The microfeatures are intended to be part of a module that can be driven by perception, language input, and memory.

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, and major divisions of history, geography, and topic. 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, characteristic lengths of events (e.g., whether events require milliseconds,

hours, or years), locations (particular cities, countries, continents, etc.) and historical times (dates or periods). 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.

Microfeatures as a Priming Context—An Example. Let us see how microfeatures could help solve the problems presented by the example from Figure 1. Figure 4 shows a partial set of microfeatures, corresponding to temporal event length or location type (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. 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 4b. 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.[9] 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 4.

Microfeatures and Context. Ideally, the particular set of microfeatures associated with a concept should serve two purposes: (1) it should be sufficient to distinguish the concept from all others, and (2) it should have shared microfeatures with all the concepts that should be associated with the given concept, but that are not related to in the ways that we would generally class as common "free associations" or n-ary relations. The set of microfeatures is thus partially definitional, but there is strictly speaking, no such thing as a complete definition for a concept in our model. In this regard the microfeatures resemble the primitives used by Wilks in his "preference semantics" (1975). This view is also similar to that expressed in Minsky (1977), that is, that concepts are defined by their positions in a network, i.e., the things to which they are connected.

We have represented all possible microfeatures as a vector, where each position of the vector corresponds to an independent microfeature, and the numerical value at that position corresponds to the level of activation of

[9] We have tried hard to be fair in contructing Figure 4a, 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 surroundings, *not* objects. All links are clearly culturally dependent though, we think, roughly in accord with current middle-class American language usage.

that feature. One might think that some microfeatures ought to be directly connected to each other by mutually inhibitory link or mutually activating links; for example "outdoor" and "indoor" microfeatures tend to be mutually inhibitory. However, both the features "indoor" and "outdoor"

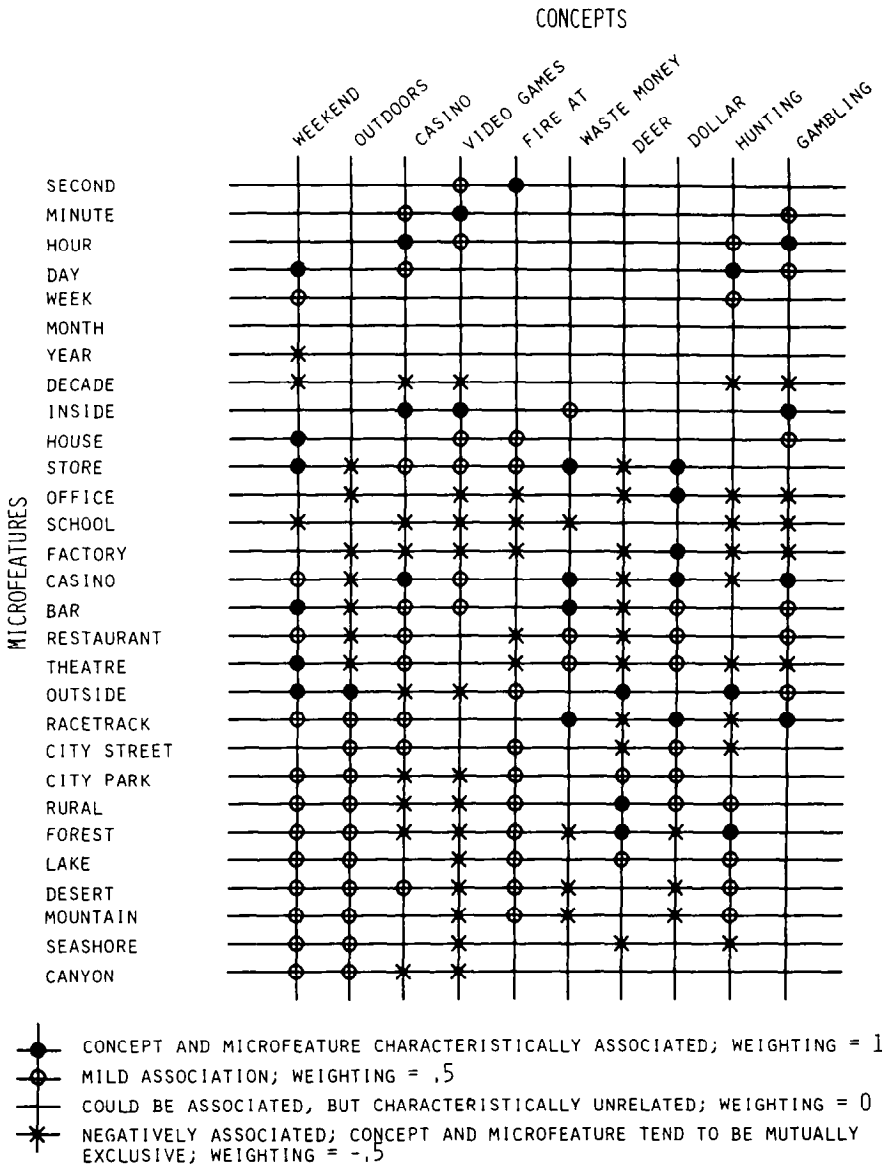

Figure 4a. This figure illustrates the use of microfeatures to provide contextual priming. At any given time, microfeatures will display some pattern of activation. Each concept has an induced activation level as a result of the microfeature activation values. The microfeature activations are modified whenever a concept is primed.

Fraction of Maximum Possible Score

Instantaneous priming effects on concepts; numbers show fraction of maximum possible score induced by the priming concepts. Microfeatures start at 0, and undergo a single priming cycle

	Primed Concepts			
Priming 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

Figure 4b. For our example, assume "weekend" is primed, with all microfeatures initially at 0. The top line of Figure 4b 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.

in the sense we wish to use microfeatures may be present in varying degrees: for example, a room with big picture windows would have a high value for indoors, but a nonzero value for outdoors, while a dense forest or space under an umbrella would have a high value for outdoors, but a nonzero value for indoors, due to perceptual enclosure and partial protection above.

Only a subset of the possible combinations of microfeature values can ever occur as contexts; although a perceptual system could in principle induce any values for the microfeatures in a full system, the real world in fact behaves in an orderly manner, so that only certain value combinations would actually be observed. Characteristic constellations of microfeature values may occur frequently or persist through time.[10] Such constellations divide the world into classes of background situations that correspond to context-setting concepts, such as "hunting," "gambling," "working in an office," or "bargaining."

The microfeature vector could be primarily driven by a memory system rather than by the perceptual system, as in vivid remembering, or in planning. While the microfeatures are primarily determined by the need to form subsets of possible situations and actions, in practice the microfea-

[10] Contexts ought to be much more persistent than individual sentence meanings, which in turn ought to be much more persistent than syntactic constructs. For example, a single noun phrase-recognizing mechanism may need to be reused several times in processing a sentence, and thus would have to be rapidly deactivated as soon as its results (e.g., a case role entry) were stored or passed to other mechanisms. For related ideas, see Woods' paper on cascaded ATN's (Woods, 1980).

tures would probably become the organizing principle of all other memories as well.

Because of memory and perceptual constraints, it would not be possible to have more than a small number of context-settings represented at once in the vector. Most typically, one context-setting at a time is represented: at least three would be needed in a situation where one is, say, planning in an imagined future, while living in the present, using a remembered context-setting for help in planning or in coping with the present. Two or three contexts would be needed for understanding my interaction with another person, one for my own world view, and one for the other person's view of the world including me. If the other person's view of me is well-enough known and sufficiently different from my own view of myself, I would need at least three contexts: (1) myself, (2) the other person, and (3) the other person's model of me.[11]

Deeper embeddings would thus of necessity be hard: e.g., "my representation of the model of the other person that the other person believes I hold." Interactions of three or more people would also be hard for us to model in this view, in keeping with the observation that with larger groups we tend not to model each individual, but to divide up the larger groups into subgroups, "us" and "them," the "good guys" and the "bad guys," "allies," who "see eye-to-eye," that is, who tend to share a world view/context, and "enemies," whose views/contexts differ (Wilks & Bien, 1983).

THE NEED FOR NEW ARCHITECTURE

Besides the evidence and phenomena which suggest a strongly integrated model for natural language, there is another reason we are looking towards a parallel model. Computer scientists, like cognitive scientists, tend to be limited by the conceptual framework of serial processing, the 30-year-old framework of the "von Neumann" machine, with its Central Processing Unit connected to its passive array of memory by a small bundle of wires.

[11] Presumably, as infants we can support only a single context, the egocentric one. Development of the ability to simultaneously support more than one context comes later, and may be the result of dividing the context vector into two subsets, each of which is reasonably complete, but which (1) can be separately activated as a whole, and (2) can support different activation patterns. Mechanisms for this could arise gradually, by processes such as reifying groups of microfeatures that frequently occur simultaneously. If we assume that activated microfeatures are sparsely scattered through the microfeature vector, it would be possible to support two or more separate contexts (e.g., me playing, mother reading). Alternatively, we may learn to "identify with" others, and use a single set of egocentric microfeatures to simulate their contexts. We hope our views may eventually have interesting connections to the ideas of Freud, Piaget, and others, but more than a footnote is premature at this point.

Backus addresses this problem in his 1977 Turing Lecture,[12] partially titled "Can Programming be Liberated from the von Neumann Style?":

> Conventional programming languages are basically high-level, complex versions of the von Neumann computer. Our thirty year old belief that there is only one kind of computer is the basis of our belief that there is only one kind of programming language, the conventional—von Neumann—language...(Backus, 1978, p. 615).
>
> Our fixation on von Neumann languages has continued the primacy of the von Neumann computer, and our dependency on it has made non-von Neumann languages uneconomical and has limited their development. The absence of full scale, effective programming styles founded on non-von Neumann principles has deprived designers of an intellectual foundation for new computer architectures (Backus, 1978, p. 616).

Backus's challenge, then, is to devise methods of computing which overcome the intellectual bottleneck in which both cognitive and computer scientists are stuck.

We have been consciously trying to answer Backus' challenge, by always informing our model with the constraints of "realizable parallelism."[13] Following in the footsteps of Fahlman, who devised a special purpose parallel processor for intersection search on a semantic network (Fahlman, 1979), and Hillis, who designed and is now building the Connection Machine (Hillis, 1981), we have designed two parallel communications architectures for modeling activation networks (Debrunner, 1983; Pollack, 1982).

CONCLUSION

We have only scratched the surface of dynamic connectionist models for language interpretation. The ideas have a long history in psychology, but a very sparse history of computer implementation. Spreading activation and lateral inhibition provide a good framework for embedding comprehension phenomena which cannot even be approached with binary serial models. We have shown that disparate knowledge sources can be smoothly integrated and can be brought to bear simultaneously on the natural language processing task. While it is clearly very crude as a cognitive model, our microfeature and concept array is a beginning toward a system which has the (correct) property of understanding an input utterance as "a 'perturbation' to an

[12] (Backus, 1978, pp. 615-616). The Turing Award is the highest honor given each year by the Association for Computing Machinery.

[13] By "realizable parallelism," we mean the communication and computation constraints from parallel computer architecture. It makes little sense to design a parallel machine in which a million processors can execute concurrently, but all have to queue up to access a central memory bank. Similarly, one cannot posit a communications network based on a tremendous crossbar switch, or use parallelism to solve an NP-hard problem in unit time!

ongoing cognitive system that is trying to make sense of things" (Winograd, 1981, p. 245). We have also shown that structural preferences such as Minimal Attachment (Frazier, 1979) can be understood as side-effects of, rather than as strategies for, a syntactic processor, and that hypotheses about lexical disambiguation in context (Seidenberg, et al., 1980; Swinney, 1979) can nicely fit into a model with lateral inhibition while 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."

Questions that still need answering include:

1. What representation system underlies (and causes) activation and inhibition links? We believe at this time that they may be based on discriminators among a distributed set of "microfeatures" of much finer grain than those used in this paper (Hinton, 1981).
2. How can a network be dynamically generated without expanding the system's power to an unreasonable degree? Obviously, a production system could be used in the manner of the READER system (Thibadeau, Just, & Carpenter, 1982) or ACT* (Anderson, 1983), but we feel that the shared "blackboard" is a bottleneck to massive parallelism (See, for example the analysis of HEARSAY II by Fennell and Lesser, 1977.) Clearly, some dynamic generation is needed to deal with the problems of embedding and crosstalk. The only approaches we know of for dynamic construction of activation networks are Feldman's Dynamic Connections (1982) and McClelland's CIP (1984, pp. 113-146), but for pragmatic reasons in our experiments we have used "normal" computer programs to generate and connect up pieces of our networks.

Generally, we are very excited about the distributed approaches to cognitive modeling which have been on the rise in the past few years, and hope that this paper contributes to their ultimate ascendency.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge: Harvard University Press.

Backus, J. (1978). Can programming be liberated from the von Neumann style? A functional style and its algebra of programs. *Communications of the ACM, 21,* 613-641.

Berwick, R. C. (1983). Transformational grammar and artificial intelligence: A contemporary view. *Cognition and Brain Theory, 6,* 383-416.

Charniak, E. (1983). Passing markers: A theory of contextual influence in language comprehension. *Cognitive Science, 7,* 171-190.

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

Church, K. W. (1980). On memory limitations in natural language processing. Master's thesis, MIT, Cambridge, MA.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82,* 407-428.

Cottrell, G. W., & Small, S. L. (1983). A connectionist scheme for modelling word sense disambiguation. *Cognition and Brain Theory, 6,* 89-120.

Crain, S., & Steedman, M. (1981, March). On not being led up the garden path: The use of context by the psychological parser. Paper presented to the Sloan Conference on Modelling Human Parsing, Austin, TX.

Debrunner, C. (1983). A two-dimensional activation cell. Working paper 41, Advanced Automation Research Group, Coordinated Science Laboratory, Urbana, IL.

DeJong, G. F., & Waltz, D. L. (1983). Understanding novel langauge. In N. Cercone (Ed.), *Computational linguistics.* Elmsford, NY: Pergamon Press.

Fahlman, S. E. (1979). NETL: A system for representing and using real-world knowledge. Cambridge: MIT Press.

Feldman, J. A. (1981). A connectionist model of visual memory. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

Feldman, J. A. (1982). Dynamic connections in neural networks. *Biological Cybernetics, 46,* 27-39.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6,* 205-254.

Fennel, R. D., & Lesser, V. R. (1977). Parallelism in AI problem-solving: A case study of HEARSAY II. *IEEE Transactions on Computers, C-26,* 98-111.

Fodor, J. (1982). *The modularity of mind.* Cambridge: MIT Press.

Ford, M., Bresnan, J., & Kaplan, R. (1982). A competence-based theory of syntactic closure. In J. Bresnan (Ed.), *The mental representation of grammatical relations.* Cambridge: MIT Press.

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

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition, 6,* 291-326.

Gazdar, G., & Pullum, G. (1982, August). *GPSG: A theoretical synopsis.* Indiana University Linguistics Club. Bloomington, IN.

Gigley, H. M. (1982, March). A computational neurolinguistic approach to processing models of sentence comprehension. (COINS Tech. Rep. 82-9). Amherst: CIS Dept., University of Massachusetts.

Goodman, B. (1984). Communication and miscommunication. Unpublished doctoral dissertation. Department of Computer Science, University of Illinois.

Hewitt, C. (1976). Viewing control structures as patterns of passing messages. (AI Memo 410). MIT AI Lab, Cambridge, MA.

Hillis, W. D. (1981). The connection machine (computer architecture for the new wave). (AI MEMO 646). MIT AI Lab, Cambridge, MA.

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

Kay, M. (1973). The MIND system. In Rustin (Ed.), *Natural language processing.* New York: Algorithmics Press.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition, 2,* 15-47.

Kwasny, S. C., & Sondheimer, N. K. (1981). Relaxation techniques for parsing ill-formed input. *American Journal of Computational Linguistics, 7,* 99-108.

MacKay, D. G. (1966). To end ambiguous sentences. *Perception and Psychophysics, 1,* 426-436.

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

Marcus, M. P. (1982). Consequences of functional deficits in a parsing model: Implications for Broca's aphasia. In M. A. Arbib, D. Caplan, & J. C. Marshall (Eds.), *Neural models of language processes*. New York: Academic.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1-72.

McClelland, J. L., & Rumelhart, D. E. (1980). An interactive activation model of the effect of context in perception. (Tech. Rep. No 91). Center for Human Information Processing, University of California at San Diego.

Milne, R. W. (1982). An explanation for minimal attachment and right association. *Proc. AAAI-82,* Carnegie-Mellon University, Pittsburgh, PA. 88-90.

Minsky, M. L. (1977, August). Plain talk about neurodevelopmental epistemology. *Proceedings of the Second International Joint Conference on Artificial Intelligence-77,* MIT, Cambridge, MA. 1083-1092.

Minsky, M. L. (1980). K-lines: A theory of memory. *Cognitive Science, 4,* 117-133.

Ortony, A., & Radin, D. (1983). Sapien: Spreading activation processor for information encoded network structures. (Tech. Rep. No. 296) Center for the Study of Reading, Champaign, IL.

Pollack, J. (1982). An activation/inhibition network VLSI cell. Working Paper 31, Advanced Automation Research Group, Coordinated Science Laboratory, Urbana, IL.

Pollack, J., & Waltz, D. (1982, August). NLP using spreading activation and lateral inhibition. *Proceedings of the 1982 Cognitive Science Conference,* Ann Arbor, MI. pp. 50-53.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *The Behavioral and Brain Sciences, 3,* 111-169.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing.* Cambridge: MIT Press.

Riesbeck, C., & Schank, R. C. (1976). Comprehension by computer: Expectation-based analysis of sentences in context. (Research Report 78), Computer Science Department, Yale University, New Haven, CT.

Sager, N. (1981). *Natural language information processing.* New York: Addison-Wesley.

Schank, R. C., Goldman, N., Rieger, C., & Riesbeck, C. (1973). MARGIE: Memory, analysis, response generation and inference in English. *Proceedings of the Second International Joint Conference on Artificial Intelligence,* Stanford University, Stanford, CA., pp. 255-262.

Seidenberg, M. S., Tanenhaus, M. K., & Leiman, J. M. (1980, March). *The time course of lexical ambiguity resolution in context.* (Tech. Rep. No. 164). Center for the Study of Reading, University of Ilinois, Urbana.

Shieber, S. M. (1983). Sentence disambiguation by a shift-reduce parsing technique. *Proceedings of 21st Association of Computational Linguistics Conference,* Cambridge, MA.

Small, S. (1980). Word expert parsing: A theory of distributed word-based natural language understanding. (Tech. Rep. No. 954). Department of Computer Science, University of Maryland.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18,* 645-659.

Thibadeau, R., Just, M. A., & Carpenter, P. A. (1982). A model of the time course and content of reading. *Cognitive Science, 6,* 157-203.

Waltz, D. L. (1978). An English language question answering system for a large relational database. *Communications of the ACM, 21,* 526-539.

Waltz, D. L. (1979). On the function of mental imagery. *The Behavioral and Brain Sciences, 2,* 567-570.

Wanner, E. (1980). The ATN and the sausage machine: Which one is baloney? *Cognition, 8,* 209-225.

Wilks, Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence, 6,* 53-74.

Wilks, Y., & Bien, J. (1983). Beliefs, points of view, and multiple environments. *Cognitive Science, 7,* 95-119.

Winograd, T. (1972). *Understanding natural language.* New York: Academic.

Winograd, T. (1981). What does it mean to understand language? In D. Norman (Ed.), *Perspectives on cognitive science.* Norwood, NJ: Ablex.

Woods, W. A. (1970). Transition network grammars for natural language analysis. *Communications of the ACM, 13,* 591-606.

Woods, W. A. (1980). Cascaded ATNs. *Journal of Association for Computational Linguistics, 6,* 1-12.