Neurally motivated constraints on the working memory capacity of a production system for parallel processing: Implications of a connectionist model based on temporal synchrony*

Lokendra Shastri
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
shastri@cis.upenn.edu

Abstract

The production system formulation plays an important role in models of cognition. However, there do not exist neurally plausible realizations of production systems that can support fast and automatic processing of productions involving variables and n-ary relations. In this paper we show that the neurally plausible model for rapid reasoning over facts and rules involving n-ary predicates and variables proposed by Ajjanagadde and Shastri can be interpreted as such a production system. This interpretation is significant because it suggests neurally motivated constraints on the capacity of the working memory of a production system capable of fast parallel processing. It shows that a large number of rules — even those containing variables — may fire in parallel and a large number of facts may reside in the working memory, provided no predicate is instantiated more than a small number of times (≈ 3) and the number of distinct *entities* referenced by the facts in the working memory remains small (≈ 10).

Introduction

Understanding language is a complex task and involves, among other things, recognizing words, accessing lexical items, disambiguating word senses, parsing, and carrying out inferences to establish referential and causal coherence, recognize speaker's plans and make predictions.[1] Nevertheless we can understand written language at the rate of *several hundred words per minute* (Carpenter & Just 1977). In view of the complexity of the language understanding task, the rapid rate at which we can understand language has strong implications and poses a challenge to computational models of cognition. In particular, it suggests that certain kinds of inferences can be drawn within a few hundred milliseconds and significant syntactic processing can occur in a similar time frame. The speed and spontaneity with which we understand language also highlights our ability to perform a class of inferences automatically and without conscious effort — as though they are a *reflex* response of our cognitive apparatus. In view of this we have described such reasoning as *reflexive* (Shastri 1990).[2]

Motivated by a concern for explaining reflexive (rapid) reasoning, Ajjanagadde and Shastri have proposed a connectionist model — let us call it SHRUTI — that can encode a large body of specific *facts*, general *rules* involving *n-ary predicates* and *variables*, as well as *IS-A* relationships between concepts, and perform a range of reasoning with extreme efficiency (Shastri & Ajjanagadde 1990; Ajjanagadde & Shastri 1991; Mani & Shastri 1991). The system performs a class of inferences in time that is independent of the size of the 'knowledge base' and is only proportional to the *length* of the shortest chain of reasoning leading to the conclusion. The reasoning system solves the dynamic (variable) binding problem (Feldman 1982; Malsburg 1986) in a neurally plausible manner: It maintains and propagates variable bindings using temporally synchronous firing of appropriate nodes. This computational model has also been used by Henderson (1991) to design a parser for English. The parser's speed is independent of the size of the lexicon and the grammar, and it offers a natural explanation for certain center embedding phenomena.

In this paper we interpret SHRUTI as a production system and examine the functional properties of the resulting production system. Such an interpretation is motivated by several factors. First, it leads to a production system with novel and interesting working memory characteristics. Second, it points the way to a neurally plausible realization of production systems. Third, it helps relate the working memory *capacity* of such a system and the time taken by each production cycle, to basic biological parameters. The interpretation also helps specify the syntactic properties of productions that can participate in reflexive processing. This aspect, however, is not discussed in this paper. The interested reader may refer to (Shastri & Ajjanagadde 1990; Shastri 1992).

A number of cognitive models are based on the

*This work was supported by NSF grant IRI 88-05465 and ARO grant DAAL 03-89-C-0031.

[1] Empirical data suggests that inferences required to establish referential and causal coherence occur rapidly and automatically during text understanding (see e.g., McKoon & Ratcliff 1980; McKoon & Ratcliff 1981; Keenan, Baillet, and Brown 1984). The evidence for the automatic occurrence of *elaborative* or predictive inferences however, is mixed (see e.g., Kintsch 1988; Potts, Keenan, and Golding 1988).

[2] A formal characterization of reflexive reasoning in terms of time and space complexity is given in (Shastri 1992): Reflexive reasoning occurs in time that is independent of the size of the long-term knowledge base and is proportional only to the length of the shortest chain of inference leading to a conclusion. Also the number of nodes required to encode a long-term knowledge base should be at most *linear* in the size the knowledge base.

production system formalism; two of the most comprehensive being ACT* (Anderson 1983) and SOAR (Newell 1990). Neurally plausible realizations of these models, however, have not been proposed. Although several aspects of ACT* such as its use of levels of activation, weighted links and decay of activation had neural underpinnings, it had not been shown how certain critical aspects of the model could be realized in a neurally plausible manner. For example, ACT* represented productions with variables, but Anderson did not suggest a neurally plausible account of how variable bindings are propagated and matched. In his exposition of SOAR, Newell has used the time course of neural processes to estimate how long various SOAR operations should take, but he has not suggested how a SOAR-like system may be realized in a neurally plausible manner (see p. 440 Newell, 1990). Although a complete mapping of comprehensive systems such as SOAR and ACT* to a neurally plausible architecture still remains an open problem, SHRUTI does provide a a concrete basis for a neurally plausible realization of production systems. Of particular significance are the specific and biologically motivated constraints SHRUTI suggests on the capacity of the working memory of a production system capable of supporting rapid 'knowledge level' parallelism.

Other researchers have proposed connectionist production systems. However, the functional characteristics of SHRUTI when interpreted as a production system are quite distinct from these connectionist models (for a detailed discussion refer to (Shastri & Ajjanagadde 1990)). For example, DCPS the distributed connectionist production system (Touretzky & Hinton 1988) only deals with productions containing a single variable. DCPS is also serial at the knowledge level and it can only apply one rule at a time. Thus in terms of efficiency, DCPS is like a traditional (serial) production system and must deal with the combinatorics of search and the associated problem of backtracking. TPPS a production system based on the tensor product encoding (Dolan & Smolensky 1989), and Conposit a system based on relative position encoding (Barnden & Srinivas 1991), are also serial at the knowledge level. Hence these systems are inappropriate for modeling reflexive processing. A connectionist system that does support knowledge level parallelism is ROBIN (Lange & Dyer 1989). However, the variable binding mechanism incorporated by ROBIN does not lead to the sort of biologically motivated constraints on working memory suggested by SHRUTI.

A Brief Overview of SHRUTI

Refer to the schematic representation of some predicates and individual concepts shown in Fig. 1. Nodes drawn as circles are what we call **ρ-btu** nodes. These nodes have the following idealized behavior: On receiving a periodic spike train, a ρ-btu node produces a periodic spike train that is *in-phase* with the driving input. Thus oscillatory activity in a ρ-btu node can lead to synchronous activity in a ρ-btu node connected to it. We assume that ρ-btu nodes can respond in this manner as long as the period of oscillation, π, lies in the interval $[\pi_{min}, \pi_{max}]$, where π_{min} and π_{max} correspond to the highest and lowest frequencies, respectively, at which ρ-btu nodes can sustain oscillatory activity. [3]

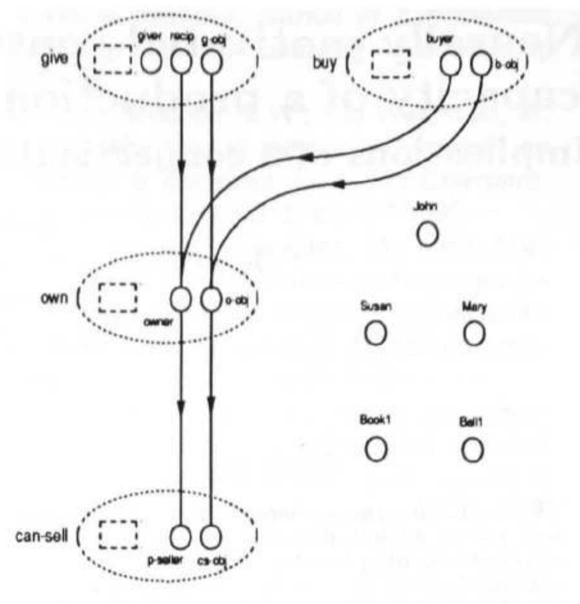

Figure 1: Encoding of predicates, individual concepts, and the rules: $\forall x, y, z\ [give(x, y, z) \Rightarrow own(y, z)]$, $\forall x, y\ [own(x, y) \Rightarrow can\text{-}sell(x, y)]$, and $\forall x, y\ [buy(x, y) \Rightarrow own(x, y)]$.

An n-ary predicate is represented by a cluster of n ρ-btu nodes (the rectangular 'nodes' shown in Fig. 1 are not relevant to our discussion). Nodes such as $John$ and $Mary$ are also ρ-btu nodes and correspond to *focal* nodes of the complete representations of the individuals 'John' and 'Mary' (Shastri 1988; Feldman 1989). A rule is encoded by linking the arguments of the antecedent and consequent predicates in accordance with the correspondence between arguments specified in the rule. For example, the rule $give(x, y, z) \Rightarrow own(y, z)$ is encoded by connecting the arguments *recip* and *g-obj* of *give* to the arguments *owner* and *o-obj* of *own*, respectively.

SHRUTI represents dynamic bindings using *synchronous* — i.e., *in-phase* — firing of the appropriate argument and concept nodes. With reference to the nodes in Fig. 1, the dynamic representation of the bindings *(giver=John, recip=Mary, g-obj=Book1)* (i.e., the *dynamic* fact $give(John, Mary, Book1)$) will be represented by the *rhythmic* pattern of activity shown in Fig. 2. Observe that while *John*, *Mary* and *Book1* are firing in distinct phases, *giver* is firing in synchrony with *John*, *recip* in synchrony with *Mary*, and *g-obj* in synchrony with *Book1*.

By virtue of the interconnections between argument nodes of the predicates *give*, *own*, and *can-sell*, the state of activation described by the rhythmic pattern shown in Fig. 2 will lead to the rhythmic activation pattern shown in Fig. 3, where the firing pattern of nodes corresponds to the dynamic bindings *(giver=John, recip=Mary, g-*

[3] We can generalize the behavior of a ρ-btu node to account for weighted links by assuming that a node will fire if and only if the weighted sum of synchronous inputs is greater than or equal to n.

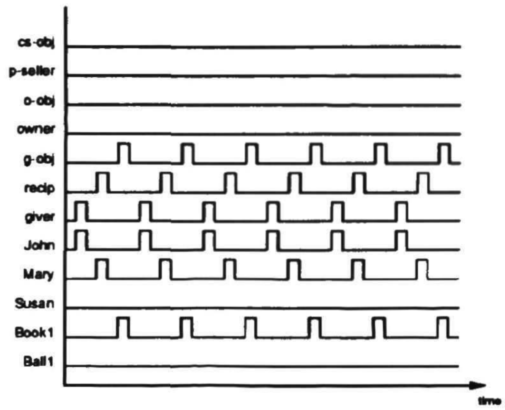

Figure 2: Rhythmic pattern of activation representing the dynamic bindings (*giver* = *John*, *recipient* = *Mary*, *give-object* = *Book*1).

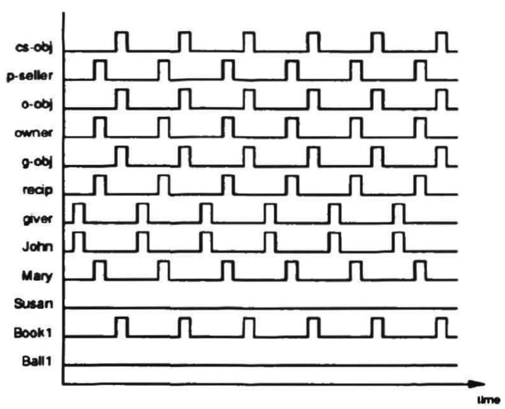

Figure 3: Pattern of activation representing the dynamic bindings (*giver* = *John*, *recipient* = *Mary*, *give-object* = *Book*1, *owner* = *Mary*, *own-object* = *Book*1, *potential-seller* = *Mary*, *can-sell-object* = *Book*1).

obj=Book1, owner=Mary, o-obj=Book1, p-seller=Mary, cs-obj=Book1) which encode the dynamic facts *give(John,Mary,Book1)*, *own(Mary,Book1)*, and *can-sell(Mary,Book1)*. In other words, given 'John gave Mary Book1', the network has inferred 'Mary owns Book1' and 'Mary can sell Book1' by using the 'rules' $give(x,y,z) \Rightarrow own(y,z)$ and $own(u,w) \Rightarrow can\text{-}sell(u,w)$.

Observe that the (multiple) bindings between *Mary* and the arguments *recip*, *owner*, and *p-seller* are represented by these argument nodes firing in-phase with *Mary*.

Conceptually, the rule application process corresponds to a ***parallel*** breadth-first traversal of a directed ***inferential dependency graph*** and a large number of rules may fire in parallel. In general, the time taken to generate a chain of inference is independent of the total number of rules and facts and is just equal to $l*\pi*\alpha$ where l equals the *length* of the chain of inference, π equals the period of oscillation of the nodes, and α is the number of cycles required for a ρ-btu node to synchronize with a connected ρ-node.

The system allows a large number of bindings — and hence, dynamic facts — to be represented simultaneously, and it also allows a large number of rules to fire simultaneously. The number of distinct entities involved in simultaneous dynamic bindings, however, is bounded by the ratio π_{max}/ω, where π_{max} is the period corresponding to the *lowest* frequency at which ρ-btu nodes can sustain and propagate oscillations, and ω is the width of the window of synchrony (i.e., two nodes firing with a lead or lag of $\omega/2$ can be considered to be firing in synchrony).

Other Representational Aspects of SHRUTI

SHRUTI can also encode *long-term* facts. The encoding of a long-term fact encodes the static bindings pertaining to the fact and rapidly recognizes that the static bindings it encodes, match the dynamic bindings existing in the system's state of activation. Given that SHRUTI represents dynamic bindings as temporal patterns, the encoding of a long-term fact behaves like a ***temporal pattern matcher*** and is described in (Shastri & Ajjanagadde 1990). With the encoding of long-term facts, SHRUTI can answer queries that follow from the encoded long-term facts and rules.

The network in Fig. 1 can only represent one dynamic instance per predicate and concept. The encoding may be extended to represent a bounded number of instantiations of each predicate and concept (for details see (Mani & Shastri 1992)). This allows SHRUTI to deal with 'bounded recursion'. However, a significant cost has to be paid for encoding multiple instantiations and as discussed below, this has implications on the working memory capacity of the associated production system.

SHRUTI can be combined with the representation of a type (IS-A) hierarchy (Mani & Shastri 1991). Such an integration allows the occurrence of types (categories) as well as instances in rules, facts, and queries. The resulting system can combine rule-based reasoning with type inheritance. For example, the system can infer 'Tweety is scared of Sylvester', based on the generic fact 'Cats prey on birds', the rule 'If x preys on y then y is scared of x' and the *IS-A* relations 'Sylvester is a Cat', 'Tweety is a Canary', and 'Canaries are birds'. The integrated system can also use type information to specify restrictions on the types of argument fillers and encode context sensitive rules such as: $\forall x : animate, y : solid\text{-}obj \;\; walk\text{-}into(x,y) \Rightarrow hurt(x)$. This rule will fire only if the first and second arguments of 'walk-into' are bound to fillers of the type 'animate' and 'solid-object', respectively.

Finally, SHRUTI can also encode rules involving multiple antecedents, thus it can encode a rule such as $\forall x,y,z \;\; love(x,y) \wedge love(y,z) \wedge notequal(x,z) \Rightarrow jealous(x,z)$.

Biological Plausibility of SHRUTI and Neurally Plausible Parameter Values

The potential of synchronous oscillation in neural representation has long been recognized (Hebb 1949; Freeman 1981; Malsburg 1981; Sejnowski 1981; Abeles 1982; Damasio 1989). Compelling evidence of the existence of synchronous activity in the brain comes from recent findings of stimulus-related synchronous oscillations in the cat visual cortex (Eckhorn et al. 1988; Gray et al. 1989). The data suggests that synchronous and rhythmic activity occurs in the brain and the time course of such activity is consistent with the requirements of reflexive reasoning. We summarize some relevant aspects of the data: i) Synchronous oscillations have been observed in the frequency range of 35 – 80 Hz (Eckhorn et al. 1988) and 35 – 65 Hz (Gray et al. 1991). Thus the observed oscillatory activity has periods ranging from about 12 to 28 msecs.; ii) Synchronization of neural activity can occur within a few (sometimes even one) periods of oscillations (Gray et al. 1991); iii) In a large number of cases synchronization occurs with precise phase-locking (zero time lag) and in most cases it occurs with a lag or lead of less than 3 msec. (Gray et al. 1991); and iv) Once achieved, synchrony may last several hundred msecs. (Gray et al. 1991).

The above data provides a basis for making coarse estimates of neurally plausible values of some of SHRUTI's parameters. The data indicates that a plausible estimate of the maximum period of oscillation, π_{max}, may be 28 msecs. and a conservative estimate of ω, the width of the window of synchrony, may be 6 msecs.

SHRUTI as a Production System

As may be evident, there exists a correspondence between SHRUTI and the production system formulation. The correspondence for the declarative memory and the production memory of a production system is straightforward: the declarative memory corresponds to the collection of long-term facts and the production memory corresponds to the collection of rules encoded in SHRUTI (each rule is a production).

Observe that dynamic bindings, and hence, dynamic (active) facts are represented in SHRUTI as a rhythmic pattern of activity over nodes in the network. In functional terms, this transient state of activation temporarily holds information during an episode of reflexive reasoning and corresponds to SHRUTI's *working memory*: a production fires if its antecedents match the contents of the working memory and introduces facts into the working memory. Observe that SHRUTI is a parallel production system that allows a large number of rules — including rules with variables — to fire in parallel as long as the capacity of the working memory is not exceeded (explained below). Furthermore, the time taken to compute a result is independent of the size of the declarative and production memory, and only depends upon the length of the sequence of productions required to produce the result.

Functional Characteristics of the Production System Implied by SHRUTI

Estimates of the working memory capacity of production system models range from very small (about 7 elements) to essentially unconstrained. SHRUTI predicts that the capacity of the *working memory underlying reflexive reasoning* (WMRR) is very large, *but* constrained in critical ways. The number of dynamic facts that can be present in the working memory at any given time is k_2p, where k_2 is a system parameter (see below) and p is the number of predicates represented in the system. Thus the number of dynamic facts that may potentially be present in the working memory is very high. But as discussed below, there exist constraints that limit the number of dynamic facts that may *actually* be present in the working memory at any given time.

Before moving on, let us clarify that the dynamic facts represented in the WMRR during an episode of reflexive reasoning should not be confused with the small number of short-term facts an agent may *overtly* keep track of during *reflective* processing and problem solving. In particular, the WMRR should not be confused with the short-term memory implicated in various memory span tasks (Baddeley 1986).

A Bound on the Number of Distinct Entities Referenced in the Working Memory During an episode of reflexive reasoning, each entity involved in dynamic bindings occupies a distinct phase in the rhythmic pattern of activity. Hence the number of distinct entities[4] that can occur as argument-fillers in the dynamic facts represented in the working memory cannot exceed $\lfloor \pi_{max}/\omega \rfloor$, where π_{max} is the maximum period at which ρ-btu nodes can sustain oscillations and ω equals the width of the window of synchrony.

As pointed out above, a neurally plausible value of π_{max} is about 28 and a conservative estimate of ω is around 6. This suggests that as long as the number of distinct entities referenced by the dynamic facts in the working memory is five or less, there will essentially be no cross-talk among the dynamic facts. If more entities occur as argument-fillers in dynamic facts, the window of synchrony ω would have to shrink appropriately in order to accommodate all the entities. For example, the value of ω would have to shrink to 4 msecs. in order to accommodate 7 entities. As ω shrinks, the possibility of cross-talk between dynamic bindings would increase until eventually, the cross-talk would become excessive and disrupt the system's ability to perform systematic reasoning. The exact bound on the number of distinct entities that may fill arguments in dynamic facts would depend on the smallest feasible value of ω. Given the noise and variation indicated by the data on synchronous activity, it appears unlikely that ω can be less than 3 msecs. Hence we predict that a neurally plausible *upper bound* on the number of distinct entities that can occur in the dynamic facts represented in the working memory is about 10.

It is remarkable that the bound on the number of entities that may be referenced by the dynamic facts in the working memory relates so well to 7 ± 2, the robust measure of short-term memory capacity (Miller 1956). This unexpected coincidence suggests that temporal synchrony may also underlie other short-term and dynamic representations.

In a large system made up of several SHRUTI-like modules, the bounds on the number of distinct entities referenced by the working memory of one mod-

[4] Note that 'Tweety', 'Tweety the Canary', 'Tweety the bird', and 'Tweety the animal' may be active simultaneously and all these count as only *one* entity.

ule is independent of similar bounds on the working memories of other modules. As we discuss in (Shastri & Ajjanagadde 1990), dynamic structures in the working memory of other subsystems may refer to different sets of entities using phase distributions local to those subsystems. Aaronson (1991) has described a connectionist interface that allows two SHRUTI-like modules, each with its own phase structure, to exchange binding information in a consistent and rapid manner.

A Bound on the Multiple Instantiation of Predicates The capacity of the working memory is also limited by the constraint that it may only contain a bounded number of dynamic facts pertaining to each predicate. This constraint follows directly from the limitation that each predicate can only be instantiated a bounded number (k_2) times. The cost of maintaining multiple instantiations of a predicate is significant in terms of space and time. The number of nodes required to represent a predicate and associated long-term facts is proportional to k_2 while the number of nodes required to encode a rule for backward reasoning is proportional to the square of k_2.[5] Thus a system that can represent three dynamic instantiations of each predicate will have anywhere from three to nine times as many nodes as a system that can only represent one instantiation per predicate. Furthermore, the worst case time required for propagating multiple instantiations of a predicate also increases by a factor of k_2 . In view of the additional space and time costs associated with multiple instantiation, and given the necessity of keeping these resources within bounds in the context of reflexive reasoning, we predict that the value of k_2 is quite small, perhaps no more than 3.

Bound on the Number of Rule Firings SHRUTI implies a production system in which any number of rules — even those containing variables — may fire in parallel as long as no relation (predicate) is instantiated more than k_2 times (where k_2 is ≈ 3) and the number of distinct entities referenced by the active facts in the working memory remains less than $\lfloor \pi_{max}/\omega \rfloor$ (≈ 10). This may be compared with Newell's suggestion (1980) that while productions without variables can be executed in parallel, productions with variables may have to be executed in a serial fashion.

Some Typical Retrieval and Inference Timings If the values of appropriate system parameters are set to the neurally plausible values identified in Section 3.1, SHRUTI performs systematic reasoning within a few hundred milliseconds. Note that we are only referring to the time taken by the internal (reflexive) reasoning process, and not considering the time taken by other perceptual, linguistic and motor processes.[6]

We choose π to be 20 msecs., assume that ρ-btu nodes can synchronize within two periods of oscillations (i.e., α equals 2), and pick the bound on the maximum number of instantiations per predicate to be 3 (i.e., k_2 equal to 3). The system takes 320 msecs. to infer 'John is jealous of Tom' after being given the dynamic facts 'John loves Susan' and 'Susan loves Tom' (this involves the production 'if x loves y and y loves z then x is jealous of z). The system takes 260 msecs. to infer 'John is a sibling of Jack' given 'Jack is a sibling of John' (this involves the production 'if x is a sibling of y then y is a sibling of x). Similarly, the system takes 320 msecs. to infer 'Susan *owns* a car' after its internal state is initialized to represent 'Susan *bought* a Rolls-Royce' (using the production 'if x buys y then x owns y' and the *IS-A* relation, 'Rolls-Royce is a car').

If SHRUTI's declarative memory includes 'John bought a Rolls-Royce', SHRUTI will take 140 msecs., 420 msecs., and 740 msecs., respectively, to answer 'yes' to the queries 'Did John buy a Rolls-Royce', 'Does John own a car?' and 'Can John sell a car?' (the last query also makes use of the production 'if x owns y then x can sell y). Note that while the first query amounts to recognizing an existing long-term fact, the second and third queries also involve inferences using other productions and *IS-A* relations in SHRUTI's declarative or production memory.

The above times are independent of the sizes of the declarative or production memories and do not increase when additional productions, facts, and *IS-A* relationships are added. If anything, these times may decrease if a new rule is added as a result of chunking.

Conclusion

We have shown that the neurally plausible model for rapid reasoning over facts and rules involving n-ary predicates and variables proposed by Ajjanagadde and Shastri can be interpreted as a production system. This interpretation leads to neurally motivated constraints on the capacity of the working memory of a production system engaged in fast parallel (reflexive) processing and helps in the estimation of the time it would take to perform such processing.

References

Aaronson, J. (1991) Dynamic Fact Communication Mechanism: A Connectionist Interface. *Proceedings of the Thirteenth Conference of the Cognitive Science Society.* Lawrence Erlbaum.

Abeles, M. (1982) *Local Cortical Circuits: Studies of Brain Function* vol. 6. Springer Verlag.

Ajjanagadde, V. G. & Shastri, L. (1991). Rules and variables in neural nets. *Neural Computation*, **3**:121-134.

Anderson, J. R. (1983) *The Architecture of Cognition.* Harvard University Press.

Baddeley, A. (1986) *Working Memory.* Clarendon Press.

Barnden, J., & Srinivas, K. (1991) Encoding Techniques for Complex Information Structures in Connectionist Systems. *Connection Science*, Vol. 3, No. 3, 269-315.

[5] A detailed discussion of the relation between k_2 and the number of nodes required to encode rules appears in (Mani & Shastri 1992).

[6] The following results were obtained using the simulator for SHRUTI described in (Mani 1992).

Carpenter, P. A. & Just, M. A. (1977) Reading Comprehension as Eyes See It. In: *Cognitive Processes in Comprehension.* ed. M. A. Just & P. A. Carpenter. Lawrence Erlbaum.

Damasio, A. R. (1989) Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33, 25-62.

Dolan, C. P. & Smolensky, P. (1989) Tensor product production system: a modular architecture and representation. *Connection Science*, 1, 53–68.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., & Reitboeck, H.J. (1988) Coherent oscillations: A mechanism of feature linking in the visual cortex? Multiple electrode and correlation analysis in the cat. *Biol. Cybernet.* **60** 121-130.

Feldman, J. A. (1989) Neural Representation of Conceptual Knowledge. In *Neural Connections, Mental Computation* ed. L. Nadel, L.A. Cooper, P. Culicover, & R.M. Harnish. MIT Press.

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

Freeman, W.J. (1981) A physiological hypothesis of perception. In *Perspectives in Biology and Medicine*, 24(4), 561–592. Summer 1981.

Gray, C. M., Koenig, P., Engel, A. K., & Singer, W. (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature.* Vol. 338, 334-337.

Gray, C. M., Engel, A. K., Koenig, P., & Singer, W. (1991) Properties of Synchronous Oscillatory Neuronal Interactions in Cat Striate Cortex. In *Nonlinear Dynamics and Neural Networks*, ed. H. G. Schuster & W. Singer. Weinheim.

Hebb, D.O. (1949) *The Organization of Behavior.* Wiley.

Henderson, J. (1991) *A connectionist model of real-time syntactic parsing in bounded memory.* Dissertation proposal. Department of Computer and Information Science, University of Pennsylvania.

Keenan, J. M., Baillet, S. D., & Brown, P. (1984) The Effects of Causal Cohesion on Comprehension and Memory. *Journal of Verbal Learning and Verbal Behavior*, 23, 115-126.

Kintsch, W. (1988) The Role of Knowledge Discourse Comprehension: A Construction-Integration Model. *Psychological Review*, Vol. 95, 163-182.

Lange, T. E., & Dyer, M. G. (1989) High-level Inferencing in a Connectionist Network. *Connection Science*, Vol. 1, No. 2, 181-217.

von der Malsburg, C. (1981) The correlation theory of brain function. Internal Report 81-2. Department of Neurobiology, Max-Planck-Institute for Biophysical Chemistry, Gottingen, FRG. 1981.

von der Malsburg, C. (1986) Am I thinking assemblies? In *Brain Theory*, ed. G. Palm & A. Aertsen. Springer-Verlag.

Mani, D. R. (1992) *Using the Connectionist Rule-Based Reasoning System Simulator.* Version 12.

Mani, D. R. & Shastri, L. (1991) Combining a Connectionist Type Hierarchy with a Connectionist Rule-Based Reasoner, *Proceedings of the Thirteenth Conference of the Cognitive Science Society.* Lawrence Erlbaum.

Mani, D. R. & Shastri, L. (1992) Multiple Instantiations of Predicates in a Connectionist Rule-Based Reasoner. Technical report MS-CIS-92-05. Department of Computer and Information Science, University of Pennsylvania.

McKoon, G., & Ratcliff, R. (1980) The Comprehension Processes and Memory Structures Involved in Anaphoric Reference. *Journal of Verbal Learning and Verbal Behavior*, 19, 668-682.

McKoon, G., & Ratcliff, R. (1981) The Comprehension Processes and Memory Structures Involved in Instrumental Inference. *Journal of Verbal Learning and Verbal Behavior*, 20, 671-682.

Miller, G.A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information, *The Psychological Review*, 63(2), pp. 81-97.

Newell, A. (1990) *Unified Theories of Cognition.* Harvard University Press.

Newell A. (1980) Harpy, production systems and human cognition. In *Perception and production of fluent speech*, ed. R. Cole. Lawrence Erlbaum.

Potts, G. R., Keenan, J. M., & Golding, J. M. (1988) Assessing the Occurrence of Elaborative Inferences: Lexical Decision versus Naming. *Journal of Memory and Language*, 27, 399-415.

Sejnowski, T.J. (1981) Skeleton filters in the brain. In *Parallel models of associative memory*, ed. G.E. Hinton & J.A. Anderson. Lawrence Erlbaum.

Shastri, L. (1988) *Semantic networks : An evidential formulation and its connectionist realization*, Pitman London/ Morgan Kaufman Los Altos. 1988.

Shastri, L. (1990) Connectionism and the Computational Effectiveness of Reasoning. *Theoretical Linguistics*, Vol. 16, No. 1, 65–87, 1990.

Shastri, L. & Ajjanagadde, V. G. (1990). A connectionist representation of rules, variables and dynamic bidings. Technical Report MS-CIS-90-05, Department of Computer and Information Science, Univ. of Pennsylvania. (Revised January 1992). To appear in *Behavioral and Brain Sciences.*

Shastri, L. (1992) A computational model of tractable reasoning — taking inspiration from cognition. Proceedings of the Workshop on Tractable Reasoning, AAAI-92, San Jose, CA. To appear.

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