Learning Language via Perceptual/Motor Experiences

Michael G. Dyer
3532 Boelter Hall,
Computer Science Department,
University of California
Los Angeles, CA 90024
E-mail: Dyer@cs.ucla.edu

Valeriy I. Nenov
Division of Neurosurgery,
School of Medicine, 74-140 CHS,
University of California
Los Angeles, CA 90024-6901
E-mail: Nenov@neurosurg.medsch.ucla.edu

Abstract

We postulate that early childhood language semantics is "grounded" in perceptual/motor experiences. The DETE model has been constructed to explore this hypothesis. During learning, DETE's input consists of simulated verbal, visual and motor sequences. After learning, DETE demonstrates its language understanding via two tasks: (a) *Verbal-to-visual/motor association* -- given a verbal sequence, DETE generates the visual/motor sequence being described. (b) *Visual/motor-to-verbal association* -- given a visual/motor sequence, DETE generates a verbal sequence describing the visual/motor input. DETE's learning abilities result from a novel neural network module, called *katamic* memory. DETE is implemented as a large-scale, parallel, neural/procedural hybrid architecture, with over 1 million virtual processors executing on a 16K processor CM-2 Connection Machine.*

Language Learning in the Blobs World Task Domain

We postulate that early childhood language semantics is "grounded" in perceptual/motor experiences. While observing and interacting in the world, each child attends simultaneously to the utterances of caregivers and, over time, learns to extract their structures and meanings. This acquisition process occurs successfully, in spite of wide-ranging differences among languages.

Here we describe DETE (Nenov, 1991; Nenov and Dyer, 1992), a system constructed to explore aspects of this *language grounding* problem. DETE is a massively parallel, procedural/neural hybrid model that consists of over 1 million virtual processors, executing on a 16K processor CM-2 Connection Machine. *Interface* modules (i.e. that map simulated visual/verbal input to learning/memory subsystems) are parallel, array-processing procedures, while *core memory modules* themselves are modeled as highly structured neural networks (termed "katamic" memories) composed of novel neural elements.

DETE receives all visual/motor input via a simulated Visual Screen (VS), consisting of 64 x 64 (i.e. 4096) "pixels". On this screen there appears a sequence of scenes. Each scene contains 1 to 5 *blobs* -- i.e., mono-colored, homogeneous (and somewhat noisy) 2-D shapes, such as circles, squares, and triangles. DETE has a single, circular retina (EYE) through which it sees a given portion of the VS at any given time. DETE can also move a FINGER icon on the VS in order to touch or push a blob.

Sequences of VS frames simulate both moving and stationary blobs of different sizes, shapes, colors, locations, speeds and directions of motion. At the same time, DETE may receive sequences of commands that move its FINGER and EYE (and zoom EYE in/out). DETE's *learning task* is to associate sensory/motor sequences with concurrent verbal sequences that describe the visual input. DETE then demonstrates its comprehension, via two *performance tasks*: (1) *Verbal-to-[visual/motor] association* -- Given only verbal input, DETE must generate (i.e. "imagine") a corresponding visual/motor sequence and (2) *[Visual/motor]-to-verbal association* -- Given only visual/motor input, DETE must generate a verbal sequence describing what it sees. Figure 1 shows DETE generating images (in its "Mind's Eye") when given only verbal input. As DETE receives verbal input, it generates corresponding visual "images" (i.e., internal representations in neural memory, that are then interpreted to produce the images in Figure 1).

DETE is designed so it can learn different language subsets. To demonstrate this ability, DETE was trained on two simplified subsets of English:

* This research was supported in part by an interdisciplinary research grant from the W. M. Keck Foundation to the first author. The CM-2 Connection Machine, on which the model is implemented, was acquired through NSF equipment centers grant #BBS-87-14206 and maintained through both the W. M. Keck Foundation grant and NSF grant #DIR-90-24251. The CM-2 is managed by the UCLA Cognitive Science Research Program.

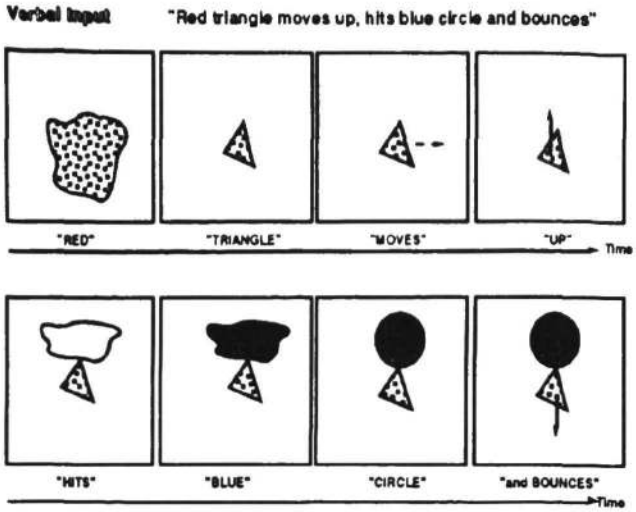

Figure 1: Performance of Verbal-to-visual association (after language learning phase). Here, "red" induces a blob with a red color; "triangle" specifies shape; "moves" activates the representation of motion with most common (default) speed and direction. Motion is shown here schematically by an arrow (which does not actually appear on the VS). "up" further specifies motion direction while "hits" induces the representation of another blob located on the motion path of the triangle. The words "blue", "circle" further refine this blob. Finally, "bounces" induces an abrupt change of motion direction of the triangular blob (indicated schematically by arrow pointing down).

FIRLAN and SECLAN. These differ in: (a) *Syntax* -- for example, in FIRLAN, blob location and motion descriptors occur *after* blob shape, while in SECLAN the order is reversed. (b) *Semantics* -- these language subsets have different vocabularies and descriptive categorizations of the perceptual world. For example, SECLAN only distinguishes position in terms of either "in_middle" or "on_periphery" while FIRLAN includes terms such as "left_of/right_of", "above/below", and "near/far".

DETE has also been taught noun/adjective gender agreement for a small, restricted subset of Spanish (e.g. "pelot*a* roj*a*" [ball red]).

Representation of Visual/Motor Input

All visual/motor input is mapped (by interface routines) to *regions of active neurons* over a set of *Feature Planes* (FPs). The 5 *visual* FPs are: Shape (SFP), siZe (ZFP), Color (CFP), Location (LFP) and Motion (MFP). Each FP is composed of a 2-D array of 16 x 16 (256) neurons. Different active regions represent different values for that feature. An *active* neuron is one that oscillates, i.e. it fires periodically (with output 1) and is silent the rest of the time (output 0). FPs have raster-linear or topographic layouts. The LFP and MFP have *topographic* layouts. If a blob is in the lower right corner of the VS, then its position is represented by a region of *active* neurons in the lower right corner of the LFP. On the MFP, the speed of a blob is represented by distance from the center, with stationary objects at the center and more rapidly moving objects toward the periphery. There are also FPs for FINGER and EYE dynamics. Figure 2 illustrates how the motions of 4 different blobs on the VS are mapped onto the Motion Feature Plane (MFP).

Motivation for Feature Planes

Feature Planes (FPs) are used as representational constructs for three reasons: (1) *Neuropsychological and Neurophysiological Support:* FPs correspond roughly to known neurophysiological and neuropsychological studies (Kandel and Schwartz, 1985) indicating both topographic mappings and that shape, position, etc. are processed in different regions of the brain and then reintegrated. (2) *Spatial Representational Analog:* Topographic layouts supply simplified, yet direct analogs for spatial features, and thus make representing space and motion easier. For example, a word like "moves" can be represented by activity anywhere away from the center of the MFP while directions of motion termed "diagonal" can be represented simply by activity anywhere in the diagonal regions of the MFP. FPs also support smooth generalization. If an object near the center of the MFP is moving slowly then objects mapped near to it will tend to be moving at about the same speed/direction. (3) *Combinatorial Learning and Generalization Capability:* Blob relationships and motions can be represented as a pattern of activity distributed over all FPs as they change sequentially in time. For example, the word "accelerate" can be represented and learned as a sequence of activations, going from the MFP's center toward its periphery. Independence of FPs also supports immediate generalization to novel combinations of known words.

Binding Features and Selective Attention

DETE can have up to 5 blobs on the Visual Screen (VS) at the same time. Multiple blobs creates a "visual binding problem" when there are separate feature planes. To avoid cross-talk, regions of activity (representing the same blob) must be "bound" across all feature planes. The solution used in DETE is one suggested in the neuroscientific literature (von der Malsburg and Singer, 1988); namely, that neurons behave as oscillators and that neurons be bound *in the temporal dimension by oscillating in phase*. All neurons in DETE oscillate at the same frequency; however, neuronal oscillations can be shifted in

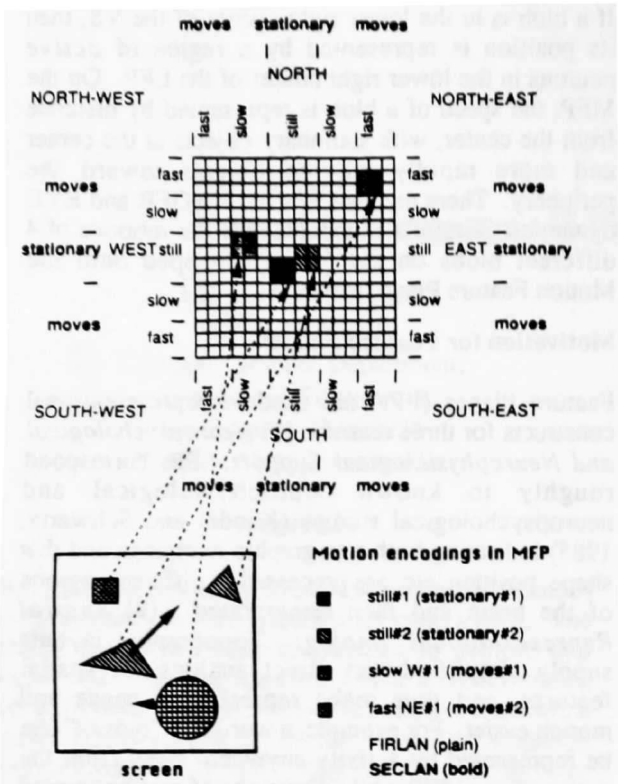

Figure 2: Motion Feature Plane (MFP). Simultaneous representation of the motions of four different objects on the Motion Feature Plane. There are two stationary blobs and another blob moving slowly to the left. The fourth blob is moving quickly in a North-East direction. The arrows do not appear on the VS but are used here simply to indicate the direction/speed of motion of blobs as their positions change over successive inputs on the VS. Illustrated also are 2 distinct motion-to-verbal mappings DETE has learned (in separate experiments): (1) In FIRLAN (i.e. words immediately surrounding the MFP), the stationary objects are described as "still" and the larger triangle as moving "fast North-East". (2) In SECLAN (words shown in **bold**), the small square and triangle are described as "stationary" while the large triangle and circle are simply described as "moves". Thus, SECLAN does not discriminate among different speeds or directions of movement.

phase. For example, if there are 3 blobs on the retina at the same time, there will be three distinct phases across all feature planes (i.e. one for each blob). In order for DETE to attend to one of these objects, there is a need for a mechanism that will focus attention on it. DETE has a (non-neural) *Selective Attention Mechanism* (SAM), which (a) assigns distinct phases to each blob on the VS and (b) makes any blob that is at the center of the EYE be in the focus of attention. DETE is designed so that only those neurons oscillating in phase with the current focus-of-attention phase can be updated. Thus,

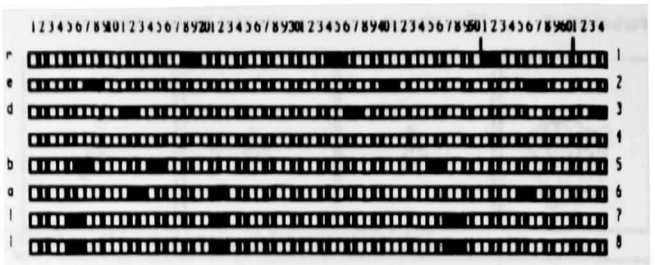

Figure 3: Gra-phonemic encoding of "red ball". For simplicity, each gra-phoneme is shown as only one 64-bit wide pattern. In DETE, each gra-phoneme is actually presented for 5 B-cycles (basic cycle of neural updating/firing), and so appears as a sequence of 5 such patterns. Pauses between words consist of 5 B-cycles of zero patterns (4th pattern here).

DETE learns about the retina-centered blob. However, the same learning mechanism allows DETE to learn also about retina-peripheral blobs, which provides a visual context.

Representation of Verbal Input/Output

Verbal input to DETE consists of a sequence of *gra-phonemes* which encode both orthographic and phonemic information. There are 26 different gra-phonemes, one for each letter (i.e. grapheme) in the English alphabet. The one-to-one correspondence between letters and gra-phonemes allows DETE to process textual input. Each gra-phoneme is represented as a pattern of active neurons over a bank of 64 "verbal' neurons. Each pattern encodes the frequencies (in Hz) of the first three formants (F1, F2, F3). Each location (loc) in the verbal bank represents a sound frequency window of 40 Hz, ranging from 270 Hz (loc 1) to 2790 Hz (loc 64). For instance, if the loc-1 neuron fires, then it represents a gra-phoneme whose first formant has an average frequency in the range of 270 to 310 Hz. Figure 3 shows how a 2-word sequence is represented.

Although gra-phonemes lack many features of real phonemes (e.g. varying duration, distinct onset vs. termination) they allow DETE to process the internal structure of words, such as inflections on verbs, and prefixes/suffixes on adjectives and adverbs (e.g. gender agreement in Spanish).

Katamic Memory

To handle dynamic tasks (such as language, vision and motion), it is important to be able to learn, recall, and recognize sequences. Also of importance are the abilities to: (a) cluster related sequences, (b) generalize to novel (yet similar) sequences, and (c)

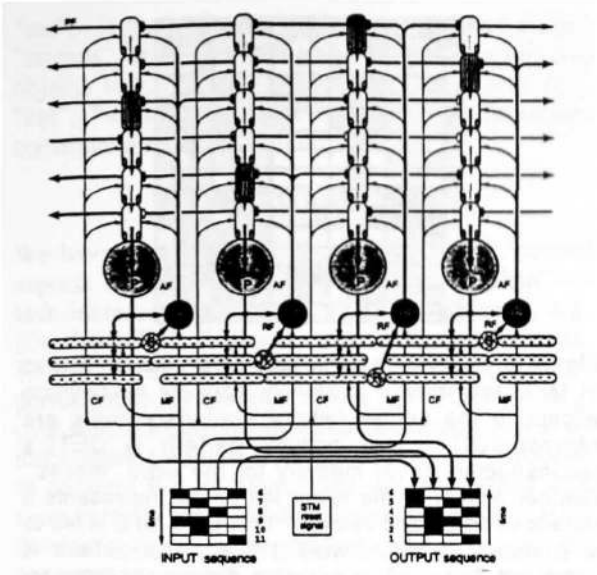

Figure 4: Katamic Memory with 4 predictrons (P), (versus 64 in verbal memory and 256 in each visual memory module in DETE model). Only 6 DCPs (vertical ovals) are shown (versus 128 in DETE model). Associated with each predictron is a recognitron (R) that samples its own pred-ictron's output and that of neighbors. Each predictron's output is passed through its recognitron's DCs and Bi-Stable Switch (BS) to connect to the DCPs of other predictrons, thus being distributing spatially (across predictrons) and temporally (across DCPs of a given, single predictron).

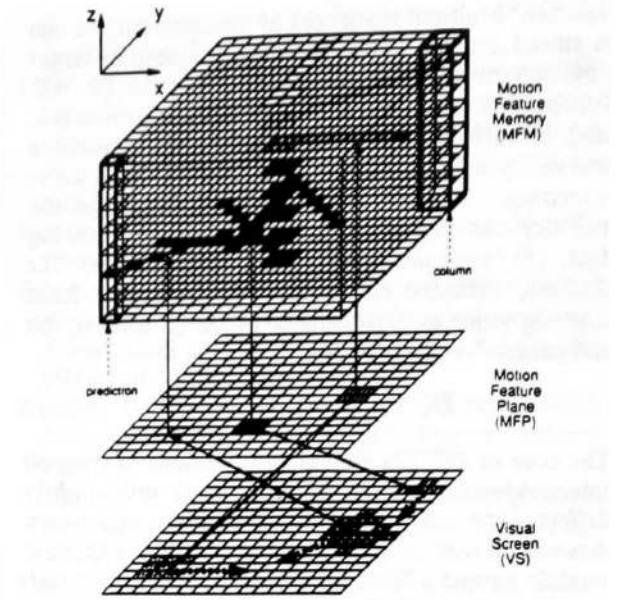

Figure 5: Spiral connectivity pattern among Motion Feature Memory (MFM) predictrons. Only 256 pixels of the VS are shown out of a total of 4096. Predictrons forming the MFM are shown as vertical bars. The DCPs of each predictron are arranged along the Z axis. Only 8 DCPs per predictron are shown here (for simplicity, recognitrons and BSSs are not shown). Each blob in the VS activates 4 units in the MFP, that then pass information to the corresponding 4 predictrons (i.e. a column) in the MFM.

recall and generate complete sequences from onlypartial cues. DETE's sequence processing abilities are based on a unique, specially designed neural network achitecture, termed "katamic memory", composed of novel, artificial neurons (Nenov, 1990). Katamic neurons have also been given special names, to distinguish them from other types of artificial neurons and neural network memory models. Each katamic memory module is constructed out of ensembles consisting of 3 neural elements: *Predictrons* (*predict*ing neu*rons*), *Recognitrons* (*recogni*tion neu*rons*) and *Bi-Stable Switches* (BSSs). Figure 4 shows a small katamic memory.

Predictrons are like real neurons in the following ways: (a) they fire a single action potential at a time (producing a stream of binary output), (b) they contain *dendritic compartments* (DCPs) (analogous to part of the dendritic tree between two branching points) that hold long-term memory information, and (c) they act as *temporal delay lines*, as the result of short-term information in the DCPs being *shifted* toward the soma (body of the neuron) over time. Thus, it takes longer for signals to reach the soma from more distant DCPs. Katamic memory is organized so that each predictron generates a 0 or 1 ateach B-cycle (basic cycle), based on the spatio-temporal relation of that output to all other bits, distributed spatially across all predictrons, and distributed temporally, across all DCPs within a predictron. The function of the *Recognitrons* and BSSs is to determine when to do internal pattern sequence completion, based on how correct were their neighboring predictrons' predictions. Katamic memory acts as a robust spatio-temporal (i.e. sequence) associator where the prediction of each next pattern is based on traces of all past histories of all previous patterns. The firing/learning B-cycle of katamic memory neurons consists of 9 steps, modeled by 16 equations; consequently, there is not enough space to discuss them here. For details, see (Nenov 1991; Nenov and Dyer 1992).

Numerous experiments on Katamic memory, reported in (Nenov 1991), show that Katamic memory has the following very useful properties: (1) *Rapid learning*. On average, only 4-6 exposures to a pattern sequence are sufficient for learning. This is 3-4 orders of magnitude improvement over recurrent PDP networks (Elman, 1990). (2) *Flexible memory*

capacity. Multiple sequences of different lengths can be stored and the model is easily scalable to larger input patterns and/or sequences of greater length. (3) *Sequence completion/recall*. A short sequence (i.e. cue) is sufficient to discriminate and retrieve a previously recorded sequence. (4) *Fault and noise tolerance*. Missing bits can be tolerated and the memory can interpolate/extrapolate from existing data. (5) *Integrated learning and performance:* The Katamic memory can switch automatically from learning mode to performance mode on a bit-by-bit and pattern-by-pattern basis.

DETE Architecture

The core of DETE's architecture consists of over 80 interconnected katamic modules, each with slightly different internal connectivity and parameter settings. Associated with each Feature Plane (FP) is a katamic module termed a "Feature Memory" (FM). For each position, e.g., in the Shape Feature Plane (SFP) there is a corresponding predictron in the Shape Feature Memory (SFM). The task of each visual FM is to encode memory traces of all sequences of activity coming in from the corresponding visual Feature Planes. How each predictron is connected to its neighboring predictrons varies with the type of feature information being encoded. For instance, in Motion Feature Memories (MFM), the predictrons are connected to one another in a *spiral* formation. Figure 5 shows this arrangement for the MFM.

Spiral connectivity, (along with phase-locking) provides a polar coordinate scheme so that the speed and direction of each blob in the VS can be uniquely represented. (Memory traces of gra-phoneme sequences are also stored via katamic memory, however, with a different connectivity configuration.)

DETE Language Learning Performance

DETE must be taught incrementally since multi-word sequences rely on single-word-to-visual associations formed earlier. First, DETE was taught the names of blobs by being given scenes of blobs with a single shape, but with varying colors, sizes, locations and motions. As a result, DETE extracts what is invariant (i.e. shape) and forms the strongest associations between the gra-phoneme sequence (e.g., "circle") and the modified DCPs of the predictrons in the Shape Feature Memory (e.g., in the region for round objects). In verbal-to-visual learning of words "circle", "square" and "triangle", DETE formed its first correct associations for each after only 4-5 presentations. To get 100% of the training data correct, however, requires 169-181 trials, mainly because irrelevant features must be varied so that the strongest associations are made for the invariant feature. Using this same approach, DETE next

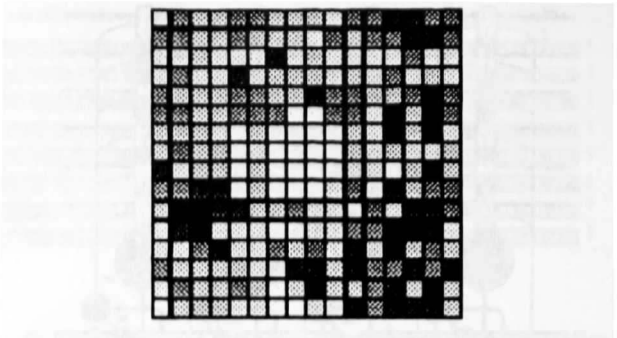

Figure 6: Activation generated on predictron somas of MFM for "moves". Activity appears everywhere except in the center (where stationary blobs are represented). This activity pattern is DETE's reconstructed visual memory for the word "moves". Greater activity in the lower left corner represents a default direction and velocity. When the MFE is fed to a winner-take-all network (WTA), the default is extracted and used to assign a direction/velocity for whatever blob (i.e. shape, color, etc.) is currently being imagined in DETE's "Mind's Eye", e.g. as in Figure 1.

learned the meanings of words for color, size, location with respect to center of the VS (e.g."above", "right", "in_center", "far", etc.). Next, DETE learned single words for actions/events -- i. e. when there is change in one or more visual features as visual frames are updated. Such words include: "moves", "accelerates", "turns","bounces", and "shrinks" (i.e. change in blob size). Once these words were learned, DETE was tested by presenting it with gra-phoneme input only, and seeing what activity appears on the corresponding FP (Figure 6). DETE has also been trained/tested on multi-word sequences (e.g. the word sequence shown in Figure 1).

Current Implementation Status and Future Directions

DETE currently runs on a 16K processor CM-2 in *Lisp with a Sun 4/830 as a front end. It uses over 1 million virtual processors (vps) and 7/8 of available heap (16K processors x .5 Mbits stack per processor)

Each of the 5 visual Feature Memories (FMs) are mapped onto a 3-D data structure with dimensions 16 x 16 predictrons (i.e. the size of each Feature Plane) x 64 (the number of per predictron) for 5 x 256 x 64 = 81,920 vps. For each FM there are 16 x 16 = 256 recognitrons. Verbal, Temporal Memories (which allow DETE to learn past/future tenses) and other modules result in a total of 1,310,720 vps.

In spite of DETE's capabilities, it has numerous limitations, including *lacking* the following: (1) verbal-to-verbal association -- needed to define abstract words in terms of known words, (2) sense of self versus others -- to learn indexicals like "you" vs.

"me", (3) goals/plans -- to learn words like "wants", "intends", and (4) model-level vision (i.e. structured objects (e.g. "chair") and composite actions (e.g. "eat", "walk"). These current, major limitations constitute directions for future research.

Conclusions

We have explored, with the DETE system, several aspects of the "Language Grounding Problem"-- a task increasingly recognized as fundamental, e.g. (Dresher, 1991; Feldman et al. 1990; Suppes et al., 1991). The superior adaptive learning and performance capabilities of the DETE system (e.g. over simple recurrent networks) are largely due to its much more complex architecture (e.g. the existence of separate feature planes; the use of neurons with more complex temporal dynamics, such as shift-delay dendritic inputs and phase-locked outputs; the connectivity within and among katamic memory modules, etc.). We believe that this use of more complex neural architecture is well justified, both computationally (i.e. to reduce what would otherwise be an enormous adaptive-learning search space) and neuroscientifically (i.e. the brain, as the result of selectional pressures over millions of years, has evolved numerous architectural structures -- each more complex than, e.g., that of a simple recurrent PDP network).

There currently remains a very large gap between the performance capabilities of *symbolic* natural language processing systems -- with their infinite generative capacity and sophisticated, human-engineered knowledge and processing constructs (e.g. capable of limited argument and belief analysis (Alvarado, 1990)), and a perceptually grounded, adaptive learning *neural* architecture like DETE. To bridge this gap will require, not only insights from computational neuroscience (Churchland and Sejnowski, 1992), but also the development of *structured* (Lange and Dyer 1989; Lange et al. 1991) and *distributed* connectionist networks specifically designed to support high-level cognition (Dyer, 1990, 1991; Miikkulainen and Dyer 1991; Sumida and Dyer, 1992).

References

Alvarado, S. J. 1990. *Understanding Editorial Text* Norwell, MA: Kluwer.

Churchland, P. S. and Sejnowski, T. J. 1992. *The Computational Brain.* Cambridge, MA: Bradford/ MIT Press.

Dresher, G. L. 1991. *Made-up Minds: A Constructivist Approach to Artificial Intelligence.* Cambridge, MA: MIT Press, .

Dyer, M. G. 1990. Distributed symbol formation and processing in connectionist networks. *Journal of Experimental and Theoretical Artificial Intelligence*, 2, 215-239.

Dyer, M. G. 1991. Symbolic neuroengineering for natural language processing: A multi-level research approach. In J. Barnden and J. Pollack (Eds.), *High-Level Connectionist Models.* (pp. 32-86). NY: Ablex Publishers.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science*, 14: 179-211.

Feldman, J. A., Lakoff, G., Stolcke, A. and Hollbach Weber, S. 1990. *Miniature Language Acquisition: A touchstone for cognitive science.* Tech. Rep. TR-90-009, ICSI, Berkeley, CA.

Kandel, E. R. and Schwartz, J. H. eds. 1985. *Principles of Neuroscience.* NY: Elsevier Science Publ. Co.

Lange, T. E. and Dyer, M. G. 1989. High-level inferencing in a connectionist network. *Connection Science*, **1**, 181-217.

Lange, T. E., Vidal, J. J. and Dyer, M. G. 1991. Artificial Neural Oscillators for Inferencing. In A. V. Holden and V. I. Kryukov, eds.. *Neurocomputers & Attention, Vol. I*, Manchester University Press.

Miikkulainen, R. and Dyer, M. G. 1991. Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science*, 15 (3), 343-399.

Nenov, V. I. 1990. Rapid Learning of Pattern Sequences: A Novel Network Model. *Proceedings of the International Neural Networks Conference.* July 15-19, Paris, France.

Nenov, V. I. 1991. *Perceptually Grounded Language Acquisition: A Neural/Procedural Hybrid Model.* Ph.D. Dissertation and Technical Report UCLA-AI-91-07, CS Dept, UCLA.

Nenov, V. I. and Dyer., M. G. 1992. Perceptually Grounded Language Acquisition: A Neural/ Procedural Model. Tech. Rep. UCLA-AI-92-03, Computer Science Department, UCLA.

Sumida, R. A. and Dyer, M. G. 1992. Propagation Filters in PDS Networks for Sequencing and Ambiguity Resolution. In J.E. Moody, S.J. Hanson and R.P. Lippmann, eds. *Advances in Neural Information Processing Systems 4*, San Mateo, CA: Morgan Kaufmann Publ., pp. 233-240.

Suppes, P., Liang, L. and Bottner, M. 1991. Complexity issues in robotic machine learning of natural language. In L. Lam and V. Naroditsky, eds. *Modeling Complex Phenomena.* NY: Springer-Verlag.

von der Malsburg, C. and Singer, W. 1988. *Princples of cortical network organization.* In: P. Rakic and W. Singer, eds. Neurobiology of Neocortex. (pp. 69-99). London: John Wiley & Sons Ltd.