A Biologically Motivated Connectionist System for Predicting the Next Word in Natural Language Sentences

João Luís Garcia Rosa

Mestrado em Sistemas de Computação – PUC-Campinas, and Mestrado em Informática – UniSantos

Rodovia D. Pedro I, km. 136 – Caixa Postal 317 – CEP 13012-970 – Campinas – SP – Brasil

joaol@ii.puc-campinas.br - Fax: +55-19-3756-7195

Abstract-- **Recent artificial neural network models lack many physiological properties of the neuron (Rocha 1992; Rosa 2001). Current learning algorithms are more oriented to computational performance than to biological credibility. The aim of this paper is to propose an artificial neural network system, called Bio-Pred, to take care of natural language processing word prediction, in a biologically inspired connectionist approach. Instead of the well-known biologically implausible back-propagation algorithm (Crick 1989; Rumelhart, Hinton, and Williams 1986), a neurophysiologically motivated one is employed (O'Reilly 1996) in a bi-directional connectionist architecture to account for next word prediction in natural language sentences. In addition, several features concerning biological plausibility are also included, for instance, distributed representations.**

Comparisons are made between Bio-Pred and a system that uses the same word representation and the same next word prediction (Rosa 2002). The differences lie in the architecture employed - bi-directional architecture versus simple recurrent network (Elman 1990) - and in the learning algorithm - a neurophysiologically inspired procedure versus the biologically implausible back-propagation. The main contribution of Bio-Pred is to make an attempt to restore biological inspiration of current connectionist systems.

Keywords: **natural language processing, biologically motivated connectionist approach.**

I. INTRODUCTION

Classical approaches to Natural Language Processing systems that account for the next word prevision problem often employ a simple recurrent connectionist architecture, with local representations of the words at the input layer (Elman 1993; Rohde and Plaut 1999). An improvement to such approach is the distributed representation for the words, adopted in Pred-DR (Rosa 2002), for the same next word prediction problem, and the same simple recurrent network structure, employing the biologically implausible back-propagation algorithm. In this paper, it is proposed a neurophysiologically inspired system called Bio-Pred, regarding both the architecture employed and the training algorithm used. Instead of the simple recurrent network, initially proposed by Elman (1990), Bio-Pred employs a bi-directional architecture. There is evidence that the cerebral cortex is connected in a bi-directional way (O'Reilly and Munakata 2000). In addition, electrical synapses are usually bi-directional (Kandel, Schwartz, and Jessell 1995). Instead of the back-propagation algorithm, Bio-Pred uses a more biological motivated one. And, at last, the computational efficiency and performance of Bio-Pred is compared to Pred-DR.

Bio-Pred attempts to predict the next word in declarative sentences presented sequentially one word at a time, giving meaning to the units of the connectionist architecture by means of distributed representations based on semantic features (Hinton, McClelland, and Rumelhart 1986; McClelland and Kawamoto 1986). This way, Bio-Pred is able to generalize to new words without increasing the number of processors in its architecture, provided that their semantic features are supplied. In addition, in a neuroscience standpoint, distributed representations seem to be predominant in the cerebral cortex (O'Reilly and Munakata 2000). The system learns to relate the input word array to its possible next word, "remembering" the previous words seen before in a semantically sound sentence. For each input word, Bio-Pred gives, as outcome, a list of probabilities of occurrence of next words in the sentence context.

II. WHY BACK-PROPAGATION IS BIOLOGICALLY IMPROBABLE?

The back-propagation algorithm is largely employed nowadays as the most computationally efficient connectionist supervised algorithm. In fact, it re-appeared in 1986 (Rumelhart, Hinton, and Williams 1986) getting the mathematical model of the neuron limitations straightened out. These limitations, regarding the linearly separable functionality of the neuron, were demonstrated by Marvin Minsky and Seymour Papert seventeen years before (Minsky and Papert 1969).

But back-propagation is argued to be biologically implausible (Crick 1989). The reason is that the supervised training algorithm is based on the error back propagation, that is, while the stimulus propagates forwardly, the error (difference between the actual and the desired outputs) propagates backwardly (figure 1). It seems that in the cerebral cortex, the stimulus that is generated when a neuron fires, crosses the axon towards its end in order to make a synapse onto another neuron input (called dendrite). Suppose that back-propagation occurs in the brain, the error must have to propagate back from the dendrite of the post-synaptic neuron to the axon and then to the dendrite of the pre-synaptic neuron. It sounds unrealistic and improbable. Researchers believe that the synaptic "weights" have to be modified in order to make learning possible, but certainly not in this way. It is expected that the weight change uses only local information in the synapse where it occurs. That is the reason why back-propagation seems to be so biologically implausible.

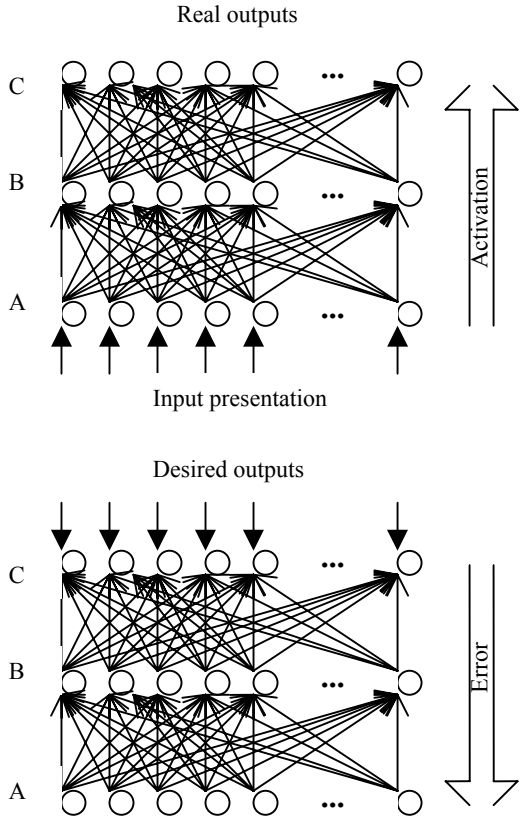


Figure 1. Schema showing the back-propagation algorithm (based on O'Reilly and Munakata 2000). The upper architecture displays the stimulus propagation phase, which consists on the presentation of the input at the input layer A, then the unidirectional propagation to the hidden layer B and then to the output layer C, generating the real output signals. So, the activation happens forwardly, in a *bottom-up* way. The lower architecture shows the error back propagation phase of the algorithm. After the propagation of the stimulus, the error, that is, the difference between real output and desired output, is propagated backwardly to the hidden and input layers, correcting the synaptic weights of the connectionist architecture. Notice that the architecture does not change: it is still unidirectional *bottom* to *up*. It is the error signal that is propagated *top-down*, not the stimulus.

It is expected that upcoming neural network models, which originally represent the connectionist computational paradigm inspired on the nervous system, restore the pioneering work, when Warren McCulloch and Walter Pitts published their paper on the mathematical modeling of the nervous cell (McCulloch and Pitts 1943). Nowadays, neural network models are considered biologically impoverished, although computationally efficient. It has been proved that neurophysiologically based systems can be computationally as effective as current connectionist systems, or even better (O'Reilly and Munakata 2000). The search for connectionist training algorithms, which are biologically plausible and computationally efficient, is the main motivation of this paper.

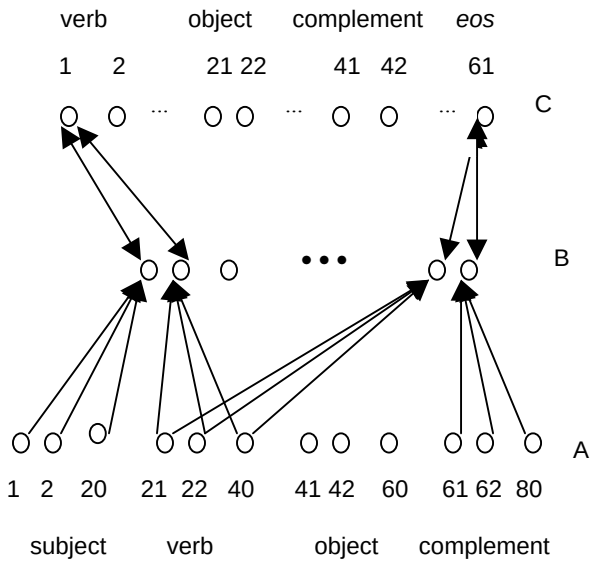


Figure 2. The three-layer bi-directional connectionist architecture of Bio-Pred. To the input layer A the words, represented by their distributed microfeature arrays, are entered sequentially, one word at a time, at their specific slot according to their syntactic category (subject, verb, object, or complement). At the output layer C, the predicted next word is shown in its specific slot also. Notice that there is no place for the subject in the output layer, since no next word could be a subject, considering the declarative sentences belonging to the training set, which is the same of Pred-DR (Rosa 2002). The final unit of the output layer (numbered 61) represents the *end of sentence marker* (*eos*).

III. The Bio-Pred System

Word prediction is considered an interesting Natural Language Processing temporal problem to be approached (Rohde 2002; Elman 1993; Rohde and Plaut 1999). In Bio-Pred, as in Pred-DR (Rosa 2002), the words of a sentence are input one at a time, at the input layer, in terms of their semantic microfeature distributed representations (McClelland and Kawamoto 1986). At the output layer, the next word (in terms of semantic features too) in the sentence context is supposed to be predicted. After checking all the distributed microfeature dimensions, the system calculates how much the actual output array is closer to a specific word. Then, the "probability" of occurrence is given, based on the distance between an average of *active* outputs and the word itself. For more details about the lexicon employed and the distributed representation adopted, see Rosa (2002).

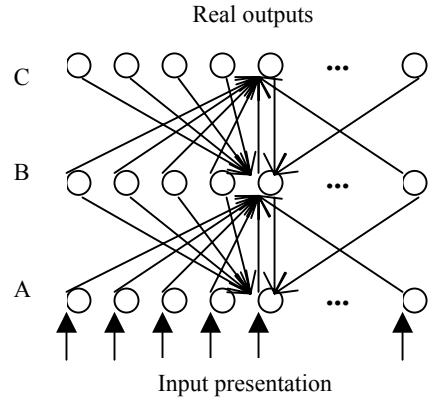
A. The connectionist architecture

The architecture employed in Bio-Pred consists of three layers, with 80 units in the input layer (to account for a four-word sentence: 20 units for each distributed representation of a word), and 61 units in the output layer, corresponding to three words of 20 units each and one unit for the *end-of-sentence marker* (figure 2). The words are represented by the semantic microfeature codification (McClelland and Kawamoto 1986), and this representation

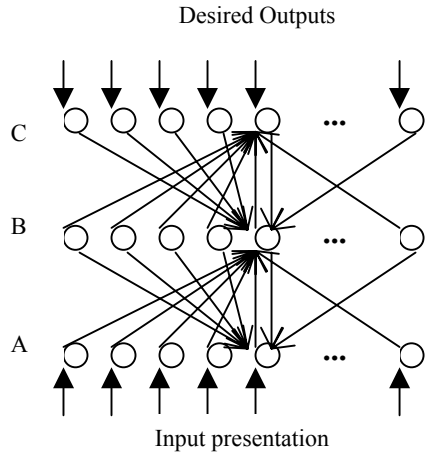
is *distributed*, in a sense that several units are used for representing one word. As a matter of fact, distributed representation is another issue considered important in a system that intends to be biologically realistic (O'Reilly 1998).

B. The training algorithm

The learning procedure is inspired by the Recirculation (Hinton and McClelland 1988) and GeneRec algorithms (O'Reilly 1996), and uses the two phases notion (minus and plus phases – figure 3).



Minus phase



Plus phase

Figure 3: The two phases of the GeneRec algorithm (O'Reilly 1996). In the *minus* phase, when the input is presented to the input layer A, there is a propagation of these stimuli to the hidden layer. Then, a hidden minus signal is generated based on the inputs and the previous output stimuli (equation 1). Then, these hidden signals propagate to the output layer C, and an actual output is obtained (equation 2). In the *plus* phase, the inputs are presented to the input layer again; there is the propagation to the hidden layer. After this, the desired outputs are presented to the output layer and propagated back to the hidden layer, and a hidden plus signal is generated (equation 3), based on the inputs and on desired outputs. Recall that the architecture is bi-directional, so it is possible for the stimuli to propagate either forwardly or backwardly.

First of all, the inputs x_i are presented to the input layer. In the minus phase, there is a propagation of these stimuli to the output through the hidden layer (bottom-up propagation). There is also a propagation of the previous actual output o_k back to the hidden layer (top-down propagation). Then, the hidden minus activation h_j^- is generated (sum of the bottom-up and top-down propagations – through the sigmoid activation function, represented by σ in equation 1). The w_{ij} represents the synaptic weights between input and hidden layers, while w_{jk} represents the weights between hidden and output layers. Finally, the current real output o_k is generated through the propagation of the hidden minus activation to the output layer (equation 2).

$$h_j^- = \sigma\left(\sum_{i=0}^{A} w_{ij}.x_i + \sum_{k=1}^{C} w_{jk}.o_k\right) \quad (1)$$

$$o_k = \sigma\left(\sum_{j=1}^{B} w_{jk}.h_j^-\right) \quad (2)$$

In the plus phase, there is a propagation from the input x_i to the hidden layer (bottom-up). After this, there is the propagation of the desired output y_k to the hidden layer (top-down). Then the hidden plus activation h_j^+ is generated, summing these two propagations (equation 3).

$$h_j^+ = \sigma\left(\sum_{i=0}^{A} w_{ij}.x_i + \sum_{k=1}^{C} w_{jk}.y_k\right) \quad (3)$$

In order to make learning possible, the synaptic weights w are updated, based on x_i, h_j^-, h_j^+, o_k, and y_k, in the way represented in equations 4 and 5. Notice the presence of the learning rate (η), considered an important variable during the experiments.

$$\Delta w_{jk} = \eta.(y_k - o_k).h_j^- \quad (4)$$

$$\Delta w_{ij} = \eta.(h_j^+ - h_j^-).x_i \quad (5)$$

C. Bio-Pred and Pred-Dr Comparisons

It was deployed two versions of Bio-Pred, Bio-Pred1 and Bio-Pred2. In both systems, the maximum acceptable error e is set to 0.02. The learning rate η is 0.25 and the hidden layer has 20 units. To Bio-Pred1 was given 24,000 training cycles, after which the system is supposed to have learned to predict the next word in declarative sentences. To Bio-Pred2, 4,057 training cycles was enough for the system to reach the error rate e. In addition to these biologically motivated systems, the comparisons included the system Pred-DR discussed earlier.

Figure 4 shows the outcomes of the systems in relation to the sentence *the wolf frightened the girl*. When the user enters *the wolf*, the system shows the next word *frighten* with probability of 81.3% in Bio-Pred1, 62.8% in Bio-Pred2, and 78.0% in Pred-DR. So, for the prediction the next word in a sentence that begins with *the wolf*, it seems that Bio-Pred1 has the greatest accuracy. Recall that a non-human animate being (*wolf*) can hit, break, and even deliver something. So, with only a word, it is very difficult to predict what will be the next word in a sentence that may occur in unconstrained contexts. When the user enters the second word (*frightened*), it is expected that a better performance would be displayed, since it is certainly easier to predict what is the next word after *the wolf frightened*, than in relation to *the wolf* only. So, Bio-Pred1 displays 82.6% for the *girl*, Bio-Pred2 shows 81.2%, and Pred-DR, 76.7%. Again, Bio-Pred1 seems to be more efficient. And finally, when *the girl* is entered, all the versions show that the next "word" should be the *end-of-sentence* marker, with 100% of certainty.

It has to be said that Bio-Pred1 is computationally more efficient than the other systems (including the non-biologically based Pred-DR) in relation to sentences belonging to the class of sentences of which *the wolf frightened the girl* is a member.

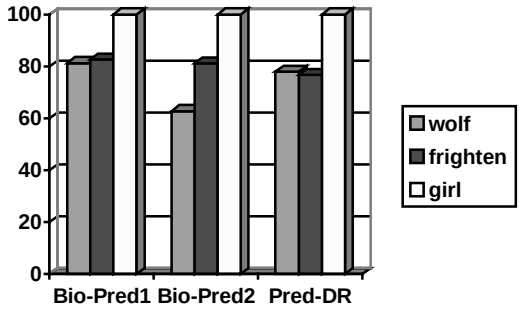


Figure 4: Responses for the sentence *The wolf frightened the girl* in two different implementations of Bio-Pred and Pred-DR. In Bio-Pred1, there are 24,000 training cycles, for a learning rate $\eta = 0.25$, 20 hidden units. For Bio-Pred2, in order to reach the error rate $e = 0.02$, it was necessary 4,057 training cycles, the same learning rate ($\eta = 0.25$), and the same hidden layer. For Pred-DR, same learning rate ($\eta = 0.25$), hidden and context layers with 20 units, and the same output error rate ($e = 0.02$), which in this case, corresponds to 2,549 training cycles. Notice that the prediction for the next word after *girl* is 100% correct, that is, the system predicts with no doubt that the *end of sentence marker* is expected in this case. This can be attributed to the fact that the verb *frighten* is normally a two operand predicate (who frightens whom).

Figure 5 shows some possible next words for *wolf* in the three presented systems. Notice that *frighten* is in the second place in all systems, after *hit* in Bio-Pred1 and Pred-DR, and after *give* in Bio-Pred2. Notice also that *deliver* and *give* have lower values, as expected, in Bio-Pred1 and in Pred-DR, but this is not true in Bio-Pred2, where *give* is the most activated word. It seems that the number of training cycles was not sufficient for the system map input word to output next word, as it was in Bio-Pred1.

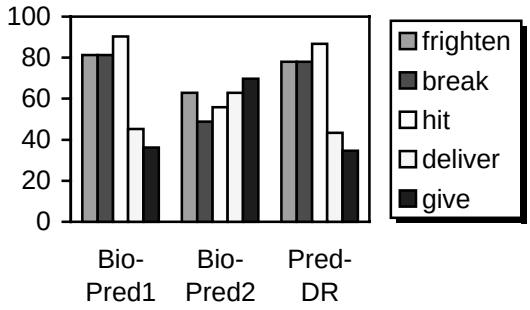


Figure 5. Responses to the input word *wolf* in the several systems. The candidates for next word are shown among several verbs.

Finally, in figure 6 it is displayed the possible next word after the phrase *the wolf frightened*. In this case, although all the three system behaviors are similar, there is a difference concerning the word *wolf*. While in Bio-Pred2 and in Pred-DR *wolf* is displayed as the most highlighted word, in Bio-Pred1 it is in the third place, after *girl* and *chicken*. So, again, Bio-Pred1 seems to have learned better the word prediction task than the other systems, including the non-biologically based Pred-DR.

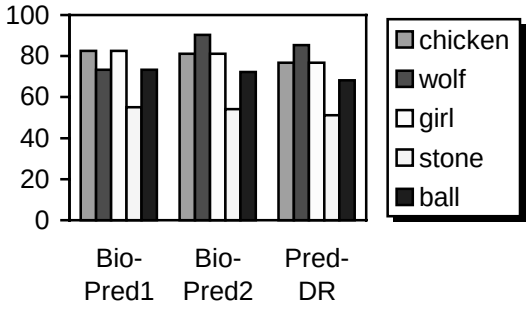


Figure 6. Responses to the input phrase *wolf-frighten* in the several systems. The candidates for next word are shown among several nouns.

The Bio-Pred versions are not so efficient with sentences like *the stone broke the vase*. The verb *break* is not so easy to process as *frighten*, since *break* may admit one, two, or three operands. When the subject is non-animate like *stone*, it is expected that *break* has one or, more often, two operands like in *the stone broke the vase*. But, certainly this necessity of decision may influence the system performance, as shown in figure 7. In this case, *break* is displayed with 66.1% of probability to be the next word

after *stone* in Bio-Pred1, 74.5% in Bio-Pred2, and 84.9% in Pred-DR.

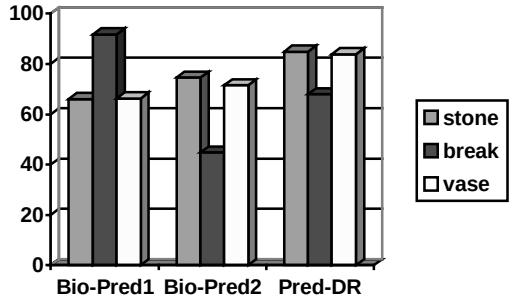


Figure 7: Responses for the sentence *The stone broke the vase* in two different implementations of Bio-Pred and Pred-DR. Notice that, in this case, the prediction for the next word after the last word *vase* is less than 100% correct, that is, the system is not so sure about the *end of sentence marker*. This can be attributed to the fact that the verb *break* can be two or three operand predicate (who breaks what or who breaks what with what).

The prediction for the second word *break* was 91.7% *vase* in Bio-Pred1, 44.8% *vase* in Bio-Pred2, and 67.9% *vase* in Pred-DR. This way, it seems that Bio-Pred1 shows better performance in relation to the phrase *the stone broke* than the other versions. Notice that because of the multi-argument possibility of verb *break*, the *end-of-sentence* marker is no longer predicted with 100% accuracy. Instead, it showed probability of 66.2% in Bio-Pred1, 71.5% in Bio-Pred2, and 83.8% in Pred-DR. It seems that, the *end-of-sentence* marker is better predicted in a non-biologically based system, at least regarding verbs with possibility of different number of arguments.

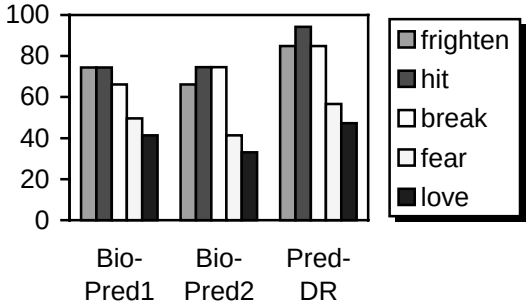


Figure 8. Responses to the input word *stone* in the several systems. The candidates for next word are shown among several verbs.

When the user enters the word *stone*, some words are highlighted as possible next words. Of course, this kind of prediction depends on what the system sees during its training step. Figure 8 shows the next word possibilities for *stone* in the three systems discussed before. Notice that *break* is expected more than other words in Bio-Pred2, while it is in the third place of Bio-Pred1 and in the second place in Pred-DR. Notice also that *fear* and *love* have lower values, as expected, mainly in Bio-Pred2. Again, it seems that a biologically based system has better performance than Pred-DR, at least in relation to the unexpected next word after *stone*.

Figure 9 shows the probable next words after the phrase *the stone broke*. Notice that *vase* is more highlighted in Bio-Pred1. As expected, *monkey* has a small percentage in Bio-Pred1 and Pred-DR. It seems, again, that the number of training cycles was not enough for the system Bio-Pred2 learn the correct relationship between the input and output words, as it seems to happen with Bio-Pred1.

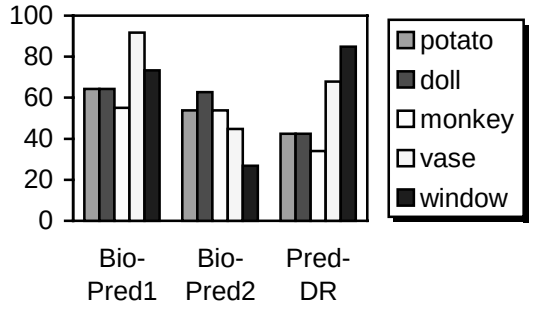


Figure 9. Responses to the input phrase *stone-broke* in the several systems. The candidates for next word are shown among several nouns.

IV. Conclusions

Bio-Pred is a connectionist natural language processing system that account for the next word prediction in natural language sentences presented one word at a time. Unlike most systems, Bio-Pred adopts a biologically motivated model, including a bi-directional architecture and a physiologically plausible learning procedure. This way, it tries to restore the neurophysiological inspiration of earlier connectionist models.

Several experiments were made to reach an architecture that, in conjunction with an error-driven task learning algorithm that resembles GeneRec (O'Reilly 1996), is able to learn the prediction of next word for sentences presented componentially one word at a time. It is important to notice that the word representation is distributed, in the sense that a set of units is used to represent one word. This is crucial in a system, which aims to be neurophysiologically based. It is presented also comparisons between Bio-Pred, deployed in two versions, and a non-biologically based system called Pred-DR. It is shown that Bio-Pred is computationally more efficient than Pred-DR, at least in relation to the training set employed.

V. References

[1] Crick, F. H. C. 1989. The recent excitement about neural networks. *Nature* 337, pp. 129-132.

[2] Elman, J. L. 1990. Finding Structure in Time. *Cognitive Science* 14, pp. 179-211.

[3] Elman, J. L. 1993. Learning and Development in Neural Networks: the Importance of Starting Small. *Cognition* 48, pp. 71-99.

[4] Hinton, G. E. and McClelland, J. L. 1988. Learning representations by recirculation, in D. Z. Anderson (Ed.), *Neural Information processing Systems, 1987*. New York: American Institute of Physics, pp. 358-366.

[5] Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. 1986. Distributed Representations. In D. E. Rumelhart and J. L. McClelland (Eds.), *Parallel Distributed Processing, Volume 1 – Foundations*. A Bradford Book, MIT Press, pp. 77-109.

[6] Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (Eds.) 1995. *Essentials of Neural Science and Behavior*. Appleton & Lange. Stamford, Connecticut.

[7] McClelland, J. L., and Kawamoto, A. H. 1986. Mechanisms of Sentence Processing: Assigning Roles to Constituents of Sentences. In J. L. McClelland and D. E. Rumelhart (Eds.), *Parallel Distributed Processing – Explorations in the Microstructure of Cognition, Volume 2 – Psychological and Biological Models*. A Bradford Book, MIT Press, pp. 272-325.

[8] McCulloch, W. S., and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, pp. 115-133.

[9] Minsky, M. L., and Papert, S. A. 1969. *Perceptrons*. Cambridge, MA. MIT Press.

[10] O'Reilly, R. C. 1996. Biologically Plausible Error-driven Learning using Local Activation Differences: The Generalized Recirculation Algorithm. *Neural Computation*, 8:5, pp. 895-938.

[11] O'Reilly, R. C. 1998. Six Principles for Biologically-Based Computational Models of Cortical Cognition, *Trends in Cognitive Science*, 2, pp. 455-462.

[12] O'Reilly, R. C., and Munakata, Y. 2000. *Computational Explorations in Cognitive Neuroscience – Understanding the Mind by Simulating the Brain*. A Bradford Book, The MIT Press, Cambridge, Massachusetts.

[13] Rocha, A. F. 1992. *Neural Nets - A Theory for Brains and Machines*, Berlin, Heidelberg: Springer-Verlag.

[14] Rohde, D. L. T. 2002. *A Connectionist Model of Sentence Comprehension and Production*. Unpublished Ph.D. Thesis. Computer Science Department. School of Computer Science. Carnegie Mellon University. Pittsburgh, PA.

[15] Rohde, D. L. T., and Plaut, D. C. 1999. Language Acquisition in the Absence of Explicit Negative Evidence: How Important is Starting Small? *Cognition* 72, pp. 67-109.

[16] Rosa, J. L. G. 2001. An Artificial Neural Network Model Based on Neuroscience: Looking Closely at the Brain. In V. Kůrková, N. C. Steele, R. Neruda, and M. Kárný (Eds.), *Artificial Neural Nets and Genetic Algorithms - Proceedings of the International Conference in Prague, Czech Republic – ICANNGA-2001*. April 22-25, Springer-Verlag, pp. 138-141.

[17] Rosa, J. L. G. 2002. Next Word Prediction in a Connectionist Distributed Representation System. In *Proceedings of the 2002 IEEE International Conference on Systems, Man and Cybernetics*. Hammamet, Tunisia, October 6-9. *Accepted for publication*.

[18] Rumelhart, D. E., Hinton, G. E. and Williams, R. J. 1986. Learning Internal Representations by Error Propagation. In D. E. Rumelhart and J. L. McClelland (Eds.), *Parallel Distributed Processing – Explorations in the Microstructure of Cognition, Volume 1 - Foundations*. A Bradford Book, MIT Press, pp. 318-362.