Toward a Connectionist Model of Symbolic Emergence

YVES CHAUVIN

PSYCHOLOGY DEPARTMENT
STANFORD UNIVERSITY

This paper examines how and why empirical results related to first-word acquisition in infants can occur in a generic associative PDP model. During learning, a network is exposed to a micro-world composed of categories made of clusters of "images" and of labels attached to these clusters. The architecture of the network allows encoding of labels and images in a common level of representation and subsequent extraction of labels from images and images from labels. If (1) the learning rule is an error-correction/steepest descent algorithm, (2) the image clusters are sufficiently "fuzzy", (3) the mapping image/label is consistent and (4) the network capacity is adapted to the size of the micro-world, this simple generic model can be shown to account for a broad spectrum of first-word acquisition data including acquisition "burst", underextensions, overextensions, gradual generalization, comprehension before production and decontextualization.

INTRODUCTION

Acquiring the meaning of words may be seen as a categorization or pattern recognition problem. The task of the infant is to classify the world into labeled categories, in agreement with the categories and labels used by adults. In this sense, there is early symbolic emergence or meaning acquisition when an entity in a modality becomes consistently mapped to another entity in a different modality. In the model below, patterns of activations are presented to a PDP network. The model has two types of inputs corresponding to two different modalities. One of them can be seen as corresponding to the auditory modality, the other to a visual modality. The model is simply exposed to a micro-world made of a micro-set of categories. Each category is composed of a set of micro-images and of an associated label. This micro-world is structured: images associated with identical labels are similar. During learning, images and labels are presented to the network, separately or together. The network learns how to build internal representations of these labels and images, and under certain conditions, to associate images with the corresponding labels.

MODEL

Microworld.

Images are simple random dot patterns constructed on a grid composed of 61 rows and 21 columns. Nine random cells are turned on to form a pattern. Before being used as input to a connectionist network, the grid is preprocessed to reduce the computational demands and create a "smearing" effect allowing a notion a similarity between patterns (Knapp and Anderson, 1984). We can call "retina" a two-dimensional layer of units (or "cells") that transform these random dot patterns into another two-dimensional pattern of activations. The units of this "retina" form a regular lattice that is superimposed on the original grid and have their receptor fields centered on a grid cell. In all the simulations described below, the shape of the receptor fields is chosen as a bi-dimensional decreasing exponential of the form $exp(-d/k)$ where k is the spread parameter and d is the distance between the center of the field and a point of the "retina".

When a pattern is presented to the model, each unit computes its activation by summing the activations due to each of the grid cells in its receptor fields. The retinal units that are too far from the active cells of the original grid cannot get enough activation and are "filtered out" of the final retinal grid. Figure 1 shows an original pattern of dots and the resulting pattern after

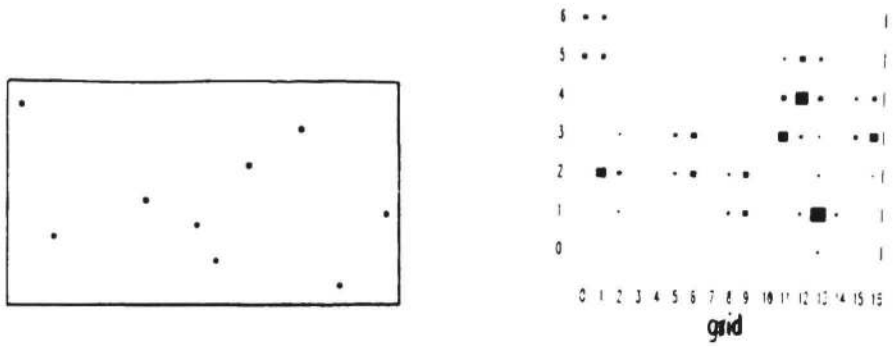

Figure 1. On the left, a random pattern of dots. On the right, a filtered pattern of dots represented on a 17x7 grid and stored into the pdp network. The size of each rectangle corresponds to the activation of the corresponding unit in the filtered grid. In this case, the filtering parameters are the following: the grain is 4, the profile is an exponential and the spread parameter is 1.2.

filtering. The filtered grid is then presented to a connectionist network for learning (see below). The microworld consists of 4 categories of such images. For each category, a basic random dot pattern was created. Out of each of these 4 basic patterns, 7 distorted patterns were generated by moving each dot around its original basic location. Each category of images is then associated to a single label (A, B, C or D).

Network Architecture.

The basic architecture of the network is shown in Figure 2. The learning rule used during the simulations is the back-propagation algorithm (Rumelhart, Hinton & Williams, 1986). The network is an auto-associative network. With this architecture, the input and output layers are identical and the network learns how to encode the incoming information in the hidden layers (Cottrell, Munro & Zipser, 1987; Zipser, 1987; Baldi & Hornik, 1988). In the present network, there are two pairs of input and output layers. One input layer theoretically corresponds to the

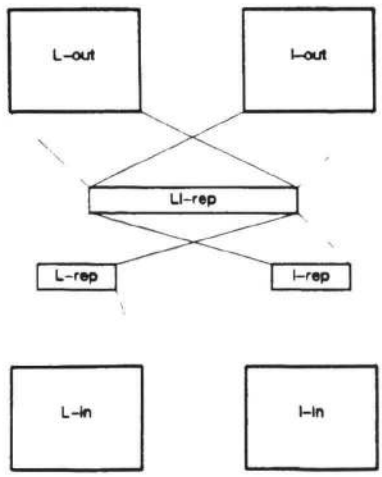

Figure 2. The network has two input layers, two corresponding output layers and three hidden layers: one for each input layer and one common to both. This last layer will encode the information that is necessary to reproduce the input patterns. Each layer is given a name that will be used in the paper. *L-in* stands for labels at the input level, *L-rep* for the representation of the labels, *L-out* for the labels at the output level, *L-in* for the "images" at the input level, *L-rep* for the representation of the "images", *L-out* for the "images" at the output level, and *LI-rep* for the representation common to both labels and "images"

"auditory modality" and the other one to the "visual modality". In the simulations, the "auditory modality" corresponds to the category labels used in the experiments. The "visual modality" will receive its input from the preprocessed random dot patterns. As we can see in Figure 2, the hidden layers *L-rep* and *I-rep* are specific to each "modality" and will specifically encode the corresponding stimuli. The common hidden layer *LI-rep* receives activations from both "modalities" and must have the capacity to regenerate the input patterns at the output level. If there is a correlation between the visual patterns and the labels, this common layer should be able to discover and represent it (Zipser, 1987).

Learning Dynamics.

Training consisted in three auto-associations: images to images, labels to labels, and images plus labels to images plus labels. With a linear auto-associative network using the delta rule, it is possible to show that the principal components of the input patterns (eigenvectors of the associated covariance matrix) are encoded "successively", in an order depending of the size of their respective eigenvalues (Chauvin, 1988). Thus, learning can be seen as a differentiation process where the "central tendencies" are encoded first. The present network is a multi-layer non-linear network using a generalization of the delta rule (back-propagation, sigmoid units). Formal analysis of this type of network have not been made possible so far. However, simulations show that, to some extent, similar processes happen during learning in both types of network.

Figure 3 represents a geometrical interpretation of the generic phenomena that happen in a simple linear network. In this case, the network is composed of two input units, 1 hidden unit, and 2 output units. The two input units correspond to two dimensions (weight and height) collected from a sample of people. The main principal component is represented in the figure by a 45 degree slanted axis. Because the considered network has only one hidden unit, only one principal component will be encoded after learning (Baldi and Hornik, 1988). This hidden unit will represent the projection of an input pattern on this major principal component. Two projections are shown in the figure. For the first one, a complete pattern is given as input to the network, corresponding to the data point x1. The activation of the hidden unit represents the projection of x1 to the major principal component. The activation of the output units represents the back-projection of this hidden unit to the original space. As we can see, the coordinates of x1 in the original space are basically retrieved by these projections. If we suppose now that only the height coordinate of x2 is given as input to the network, the coordinate becomes projected to the

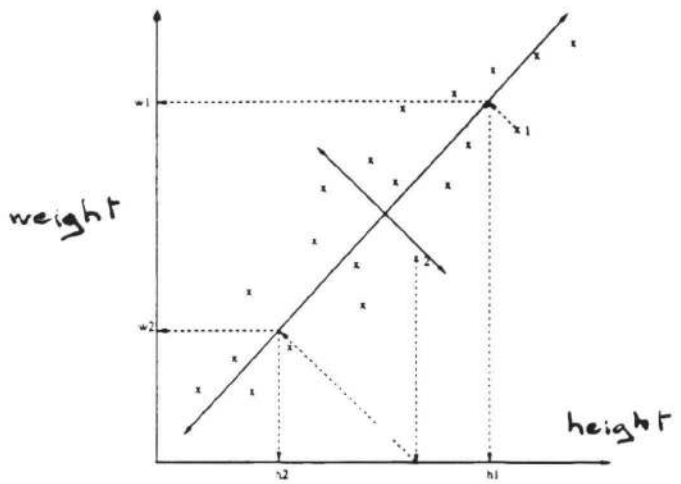

Figure 3. Geometrical interpretation of pattern completion by linear projection on the major principal component (see text). Two points are projected. The first one, x1, is projected from the complete original position to the major principal component. For x2, the "height" coordinate only is projected. Each of these principal component projections is then reinterpreted by back-projection in the original weight/height space.

major principal component and then back to the original space. As we can see, some information has been retrieved about the "weight" of x2. This corresponds to a pattern completion phenomenon by projection on the principal component.

SIMULATIONS

Categorization

After learning, a label presented to *L-in* reproduces itself in *L-out* and an image presented to *I-in* reproduces itself in *I-out*. The layers *L-rep, I-rep,* and *LI-rep* then represent the compressed information (Cottrell, Zipser, & Munro, 1987) of the input patterns. For a right amount of hidden units, presented with an image, the network is able to produce the label that corresponds to the associated category: there is *production*. Presented with a label, the network is able to give an image that basically corresponds to the average of all the images that have been stored with the same label: there is *comprehension*. For a sufficient number of images per category and a right set of low-level filtering parameters, the prototype effect can be observed for comprehension, as observed with infants (Thomson & Chapman, 1977), and production. Interestingly, because images form clusters, knowing the shape of an image provides some information about what the label should be. However, the network is not being trained to produce a label when an image is presented. The network does use the image information and automatically learns the cross association only because there is auto-association image to image and subsequent cluster extraction during learning: the internal representation of the images is *necessary* for the development of the labeling process.

Gradual Generalizations

Three levels of distortion are used to test generalization of categorization to new images (the network was trained on patterns created with the medium level only). These levels of distortions correspond to different standard deviations of a Gaussian noise added to each dot location of the prototypical images. Figure 4 shows the acquisition orders of each distortion level. As we can see, the network gradually learns how to generalize production and comprehension to more and more distorted patterns. Thomson and Chapman (1977) and others observed gradual generalization for comprehension with infants. Interestingly, generalization actually occurs earlier and faster for comprehension than for production.

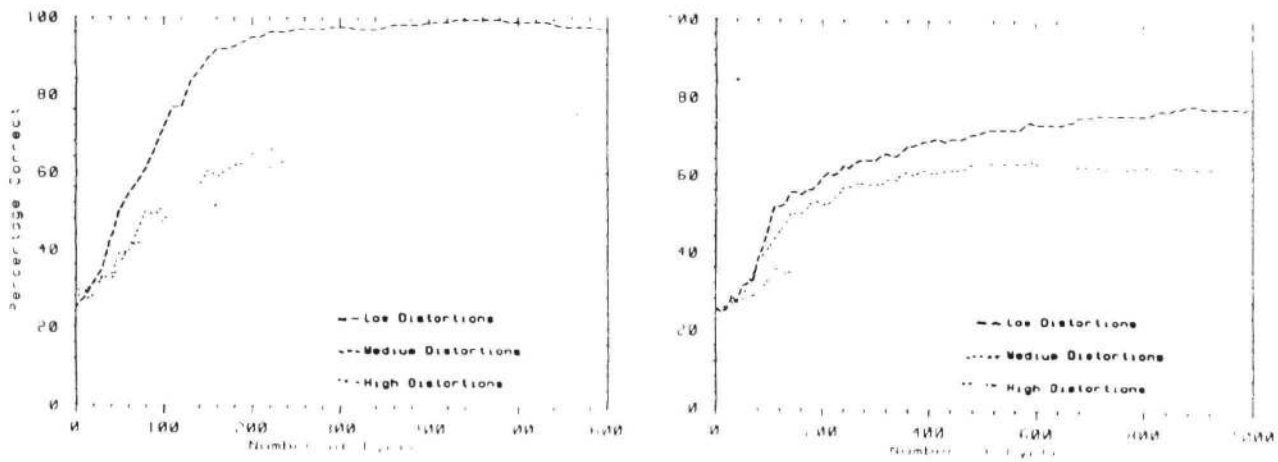

Figure 4. Gradual generalization for production and comprehension. The network gradually learns how to respond correctly to more and more distorted patterns.

Comprehension before Production.

Figure 5 shows production and comprehension data as a function of the number of learning cycles. We can see that production and comprehension performances are similar after sufficient training but that comprehension performance is clearly higher than production during early learning, in agreement with comprehension/production data observed during human first word acquisition (e.g. Bates, 1976). Label and image features can actually be considered as category features. Among these features, labels are the most "significant" because they are consistently present in all the examples of the category. For that reason, labels become good indicators of the category clusters and will allow good reconstitution of the images. In contrast, image features may be present or absent or "graded" among the examples and will not reconstitute the labels as well.

Acquisition Rates

Typically, categorization rates are low during early learning and suddenly increase as learning goes on. This initial period is usually longer for production than for comprehension and the production rate increase is not as sharp. During the differentiation process, the network actually learns how to distinguish the categories before distinguishing the exemplars within each category. As long as the categories are not distinguished, the network is still able to categorize some of the patterns correctly, "by chance", depending on their "projection" to the category averages. However, there is very little generalization during this period: the network is only able to generalize to patterns that are closely correlated to already stored patterns. When the network starts to "realize" that there exist category clusters, by "pulling apart" the corresponding averages, there is generalization and sharp increase in the acquisition rates. This sharp increase in comprehension and production rates can be compared to the well-known vocabulary explosion observed in production with humans (e.g. Barrett, 1983).

Decontextualization and Underextensions

Here, decontextualization is viewed as the process of shifting from temporarily associating a label with a single image to extending the association to the complete set of images corresponding to the label. Simulations show various cases of decontextualization depending on the initial weights of the network. In the most common case, a label is correctly mapped to only one image for some time and becomes slowly extended to the whole category while being generalized to new category examples. In another case, two category labels are decontextualized one after the

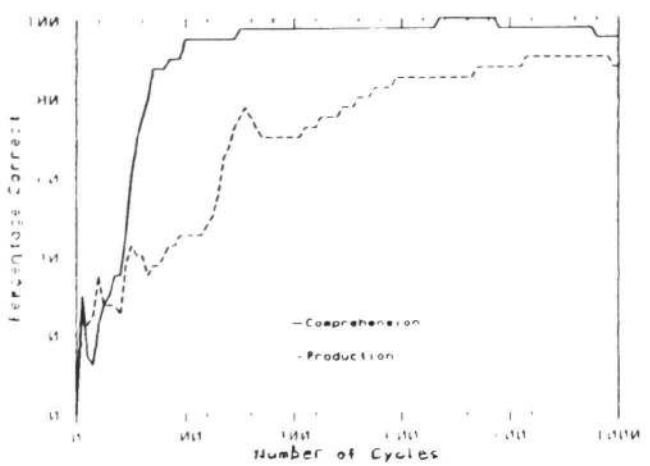

Figure 5. Comprehension versus production. During early learning, performance on comprehension is better than on production.

other: one is decontextualized much later than the other one, but also much faster. In other cases, the opposite phenomenon occurs, where a first category label is quickly decontextualized whereas another one is decontextualized much later and much more slowly. Interestingly, the simulations are very consistent with recent human data on decontextualization. There does not seem to exist an initial period where labels are first slowly decontextualized and a subsequent period where they are decontextualized from the onset (as previously suggested by Bates, 1979). In agreement with Barrett (1986, In Press), a label can be correctly mapped to a complete category early in learning while some other label might appear later and be slowly decontextualized. Furthermore, underextension followed by a forgetting stage followed by correct extension might occur, as observed by Bloom (1973).

Overextensions.

During very early learning, the network extracts the general average taken over the entire set of stored patterns. However, after this initial period, the network encodes the first principal component of the patterns and finds a steepest slope in a direction that might better correspond to one of the categories. Any other pattern correlated with this biased category will be similarly categorized by the network and overextensions might occur. In the model, if a label unit activation in *L-out* is above a given threshold but does not correspond to the label associated with a presented image, it can be considered as an over-extension of the indicated category. Simulations show that for production, over-extensions do occur for some of the categories and can be highly variable. When they occur, they are generally followed by periods of underextension, before being slowly readjusted to a correct "extension level". Overextensions are also much more likely to occur during early learning than during late learning. Figure 6 shows the total amount of overextensions and the total amount of correct extensions as a function of the number of cycles during a typical run. Overextensions also occur during comprehension. However, they start earlier, they end earlier, and they are not as numerous as overextensions during production. This difference between production and comprehension is also due to the fact that labels are good cluster indicators. Again, there are interesting similarities between the way the network learns and the way children acquire their first words. First, the network produces overextensions, in spite of equiprobable presentations (e.g. Rescorla, 1980). Second, the overextensions occur mostly during early learning: the late acquired categories are not overextended (Rescorla, 1976). Third, overextensions can be followed by a "recession" stage before correct extensions begin to take place. Fourth, overextensions are more frequent in production than in comprehension (Thomson & Chapman, 1977). Fifth, if a category is being correctly extended, then no other category can overextend to it (Leopold, 1949).

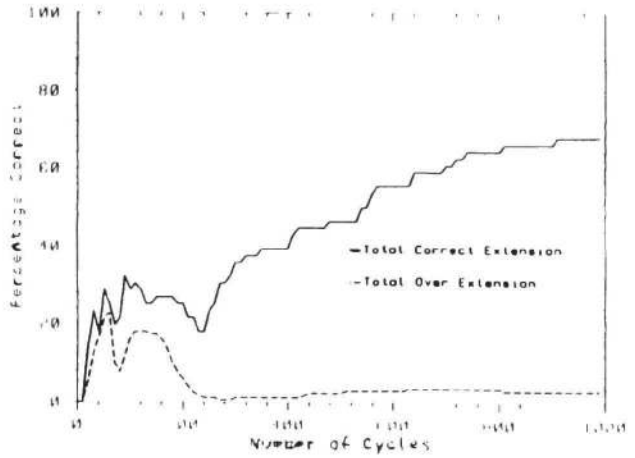

Figure 6. Total number of correct and over extensions as a function of the number of cycles for production during a typical run.

SUMMARY AND DISCUSSION

The patterns stored in the network can be viewed as multi-dimensional correlated vectors. The delta learning rule encodes these sets of vectors by first representing their major principal components. Figure 3 shows a geometrical interpretation of the phenomena occurring in a simple linear network for a low dimensional space. The projections represent the "knowledge" that the network has about the world. To retrieve the world knowledge from this representation means back-projecting these projections to the original space. If there is sufficient information compression in the hidden layers, labels will be retrieved from images and prototypical images from words (production and comprehension). The strictly consistent mapping between labels and image categories creates learning asymmetries between comprehension and production. In general, the direction of the main principal components depends on the image clusters and on the consistency between labels and image clusters. Therefore, a category prototype closer to the first principal component might dominate the whole set of examples during early learning, creating over-extensions to the related category. As the "category directions" are being discovered by the network, it becomes much easier to classify the examples belonging to the corresponding clusters (comprehension and production rate explosions). From this onset onwards, examples are really classified according to their prototypical directions. Finally, images can reproduce labels only because there is a differentiation process happening during auto-association of the images. During this process, the image clusters are reinforced in the internal representations and the labels can "understand" the information coming from the images. In this sense, the internal representation of the world, seen as a principal component or "central tendency" analysis, is necessary for the linguistic mapping.

The goal of this study is not to construct a realistic model of first word acquisition. Rather, it is to explore if phenomena related to symbolic emergence in infants could be "naturally understood" in a Parallel Distributed Processing framework. The differentiation process proposed by psychologists such as Piaget and Werner during early language acquisition is reminiscent of the phenomena occurring in simple linear networks using an error correction rule. Therefore, the original idea was to store in a network using such a rule, a set of patterns that would correspond to labels and images and to observe how and understand why associations between these labels and patterns could be built during learning. The network had to internalize the presented patterns in such a manner that resulting representations would be able to reproduce images and labels from images or labels. This constraint forced a level of representation that would compress labels and images into a common encoding layer. The present network can then be seen as a simple generic model that "embodies" these very general principles. Interestingly, the network was able to mimic quite a number of first word acquisition phenomena just by using these few principles.

Acknowledgments.

I am greateful to Dave Rumelhart and to the PDP research groups at UCSD and Stanford University for useful discussions. I am especially thankful to Yoshiro Miyata for the use of his simulator SunNet.

REFERENCES

Baldi, P., & Hornik, K. (1988). Neural networks and principal component analysis: Learning from examples without local minima. *Proceedings of the Conference on Neural Information Processing Systems, Denver, CO.*

Barrett, M. D. (1983). The course of early lexical development: A review and an interpretation. *Early child development and care, 11*, 19-32.

Barrett, M. D. (1986). Early semantic representations and early word usage. In S. A. Kucsaj, M. D. Barrett (Ed.), *The development of word meaning*. New York, NY: Springer-Verlag.

Barrett, M. D. (In Press). Early language development. In A. Slater, G. Bremer (Ed.), *Infant Development*. London: Erlbaum.

Bates, E. (1976). *Language and context: The acquisition of pragmatics*. New York, NY: Academic Press.

Bates, E., Benigni, L., Bretherton, L., Camaioni, L., & Volterra, V. (1979). *The emergence of symbols: Cognition and communication in infancy*. New York, NY: Academic Press.

Bloom, L. (1973). *One word at a time*. The Hague: Mouton.

Chauvin, Y. (1988). *Symbol Acquisition in Humans and Neural (PDP) Networks*. Unpublished Doctoral Dissertation. University of California, San Diego..

Cottrell, G. W., Munro, P., & Zipser, D. (1987). Learning internal representations from gray-scale images: An example of extensional programming. *Proceeding of the Ninth Annual Conference of the Cognitive Science Society, Seattle, WA*, 461-473.

Knapp, A. G., & Anderson, J. A. (1984). Therory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 616-638.

Leopold, W. F. (1949). *Speech development of a bilingual child: A linguist's record* (Vol. 3). Evanston, Ill: Northwestern University Press.

Rescorla, L. (1976). Concept formation in word learning. *Unpublished doctoral dissertation.* Yale University

Rescorla, L. (1980). Overextension in early language development. *Journal of Child Language, 7*, 321-335.

Rumelhart, D. E., Hinton, G. H., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland (Ed.), *Parallel distributed processing. Explorations in the microstructure of cognition. (Vol 1).* Cambridge, Ma: MIT Press/Bradford Books.

Thomson, J. R., & Chapman, R. S. (1977). Who is "Daddy"? The status of two-year-olds' over-extended words in use and comprehension. *Journal of Child Language, 4*, 359-375.

Zipser, D. (1986). *Programming networks to compute spatial functions* (Tech. Rep. No 8608). University of California, San Diego, Insitute for Cognitive Science.