Proceedings of the Eighth Annual Conference of the Cognitive Science Society
Lawrence Erlbaum Associates: Hillsdale, N. J., 1986, 72-89.

THE SPACING EFFECT ON NETTALK, A MASSIVELY-PARALLEL NETWORK

Charles R. Rosenberg
Cognitive Science Laboratory
Princeton University

Terrence J. Sejnowski
Department of Biophysics
Johns Hopkins University

ABSTRACT

NETtalk is a massively-parallel network that learns to convert English text to phonemes. In NETtalk, the memory representations are shared among many processing units, and these representations are learned by practice. In humans, distributed practice is more effective for long-term retention than massed practice, and we wondered whether learning in NETtalk had similar properties. NETtalk was tested on cued paired-associate recall using nonwords as stimuli. Retention of these target items was measured as a function of spacing, or the number of interspersed items between successive repetitions of the target. A significant advantage for spaced or distributed items was found for spacings of up to forty intervening items when tested at a retention interval of 64 items. Conversely, a significant advantage for massed items was found if testing immediately followed study. These results are strikingly similar to the results of many experiments using human subjects and suggest an explanation based on distributed representations in massively-parallel network architectures.

INTRODUCTION

CRR was supported in part by a research grant (487906) from IBM, by the Defense Advanced Research Projects Agency of the Department of Defense and by the Office of Naval Research under Contracts Nos. N00014-85-C-0456 and N00014-85-K-0465, and by the National Science Foundation under Cooperative Agreement No. DCR-8420948 and under NSF grant number IST8503968. TJS was supported by grants from the National Science Foundation, System Development Foundation, Sloan Foundation, General Electric Corporation, Exxon Education Foundation, Allied Corporation Foundation, Westinghouse, and Smith, Kline & French Laboratories.

We are indebted to Dr. T. Landauer for calling this problem to our attention. We also wish to thank George Miller and Stephen Jose Hanson for many helpful comments, Katherine Miller for expert editorial assistance, and Bell Communications Research for generously providing computational support.

In 1885, Ebbinghaus noted that "with any considerable number of repetitions a suitable distribution of them over a space of time is decidedly more advantageous than the massing of them at a single time" (Ebbinghaus, 1885/1964 p.89). Since then, the spacing effect has been found across a wide range of stimulus materials and tasks, semantic as well as perceptual/motor, and has even been found when the repetitions are across modality, or across languages, if bilinguals are employed as subjects (see Hintzman, 1974, for a review). The ubiquity of these results suggests that spacing reflects something of central importance in memory. However, despite over a hundred years of research, the spacing effect, as general as it is, continues to defy adequate, or at least, simple, explanation.

Perhaps the most popular account of the spacing effect is the encoding variability hypothesis (e.g. Melton, 1970; Martin, 1968; Glenberg, 1979). This hypothesis makes two major assumptions: (1) stimuli are encoded relative to the context, or environment, in which they occur, and (2) the probability of retrieval is positively correlated with the similarity of the context at retrieval to the context at encoding.[1] Given a continuously evolving environment, two trials that occur back-to-back will tend to share more context with each other than two trials widely separated in time.[2] Distributing practice will hence be advantageous to the extent that the two repeats of the to-be-remembered (TBR) item are encoded more independently, thus boosting the probability of the item's retrieval in a randomly chosen context presumably by increasing the number of possible retrieval routes.

Overall, the encoding variability hypothesis has found only limited empirical support (Hintzman, 1976). One recent failure of the hypothesis is the study by Postman and Knecht (1983) in which the encoding contexts of words were varied by embedding them in different sentences, either one sentence repeated three times or in three sentences each only once repeated. In the cued recall task, cueing was with one or three of the sentence frames (with the TBR word deleted). According to the encoding variability hypothesis, retrieval should be greater in the multiple context condition. Nevertheless, they found no difference in free recall of the target words, tested either immediately or after 24 hours. In fact, cued recall with a single sentence frame led to higher recall rates for targets that appeared in single contexts than targets that appeared in multiple contexts, a trend in a direction opposite to that predicted by the hypothesis.

---

[1] The notions underlying the encoding variability hypothesis were originally derived from Estes's stimulus sampling and fluctuation model (1959).

[2] The precise use of the term "context" has not always been consistent among investigators. See Maki & Hasher (1975) for a discussion and empirical investigation of this issue.

Postman and Knecht concluded that encoding items in different contexts does not necessarily improve retention, and may, in fact, lead to diminished retention. An overriding factor may be the strength of specific cue-target associations, built up by repeating the item in identical contexts. That is, many weak retrieval routes are not necessarily better than one strong one.

If we assume, as the Postman and Knecht study suggests, that the spacing effect depends to some extent upon the repetition of specific items, and not necessarily on the encoding of items relative to a continuously varying context, then the following question arises: Which repetition of the item, the first or the second, is less effectively processed or encoded when the two presentations occur back-to-back? Those theories that claim that the first presentation is deficient include the rehearsal-buffer theory (Atkinson & Shiffron, 1968; Rundus, 1971) and a version of consolidation theory (Landauer, 1969). In either case, the disadvantage found for massed practice is the result of the interruption of an ongoing encoding process by the immediate occurrence of the second item. Bjork and Allen (1970) found, however, that interposing a more difficult task between repetitions did not disrupt this encoding process, as both rehearsal-buffer theory and the consolidation hypothesis would predict. To the contrary, they found *improved* retention in the difficult task condition.[3] This result is hard to reconcile with either theory.

The other alternative is that the *second* massed presentation is more poorly processed or somehow less effective. It has been proposed that subjects habituate (e.g. Hintzman, Block, & Summers, 1973) and therefore cannot process the second massed presentation as effectively as they can the first. It is not clear how this proposal could explain the Bjork and Allen result, unless the intervening difficult task in some way releases the habituation from the first item. In addition, attempts to overhabituate to a target item by presenting the item for longer durations have been unsuccessful (Hintzman, Summers, & Block, 1975).

Another suggestion is that subjects do not attend as effectively (e.g. Shaughnessy, Zimmerman, & Underwood, 1972) to the second occurrence of an item if it closely follows an identical first item in time. But efforts to force subjects to attend to the second occurrence in various ways have indicated no reduction in the spacing effect (e.g. Hintzman, Summers, Eki, & Moore, 1975).

Jacoby (1978) has offered an account of spacing in terms of processing effort. That is, in the massed presentation condition, subjects have the first item consciously available when

[3] This result has been replicated by Tzeng (1973).

they are presented with the second item and consequently do not have to process the second item to the same degree. As a result, processing on the second massed item is not as great as that on the first and does not form as rich a code. This explanation seems to give a coherent account of all the evidence thus far mentioned: It does not depend on encoding items relative to a dynamic context, and it accounts for the Bjork and Allen results, since interposing a difficult task could plausibly make the second item less available, requiring more processing. However, it leaves unclear what "processing effort" (not to mention "consciousness") involves.

All these theories attempt to explain spacing in terms of concepts such as encoding, habituation, and consolidation, which make little reference to the actual form of the memory representation, although implicit in some of the explanations is the assumption that individual items have local representations. Another approach is to seek an explanation at the level of the representation: *It may matter how the information is stored in the system.* One way to explore this possibility is to construct explicit models that incorporate particular memory representations and learning mechanisms and to test them with the same experimental paradigms that have been used to study human memory.

It will be demonstrated that the spacing effect is a natural consequence of learning in a network with distributed memory representations and an incremental learning procedure. In this framework, learning consists of modifying connections in the network so that this information is retained as accurately as possible within the constraints imposed by the number of available connections. This way of storing information is fundamentally different from a local representation where individual items can be stored independently of one another, as in a computer memory. In a distributed representation a single connection can participate in the storage of many items, and conversely a single item is stored in many connections. Approaching memory in this way has already led to new insights in the domains of categorization and concept formation (McClelland & Rumelhart, 1985; Anderson, Silverstein, Ritz, & Jones, 1977; Eich, 1982; Amari, 1977; Kohonen, Oja, & Lehtio, 1981). These models of memory are inspired by the parallel architecture of the brain (Ballard, Hinton, & Sejnowski, 1983; Feldman & Ballard, 1982).

NETtalk

We have recently described NETtalk (Sejnowski & Rosenberg, 1986), a massively-parallel network that learns to translate letters in English text into phonemes and associated word stress. It achieves approximately 95% accuracy per letter without access to information about semantics or syntax. In NETtalk, the learning occurs by modifying the strengths of connections between a large number of simple and identical processors, or units. These

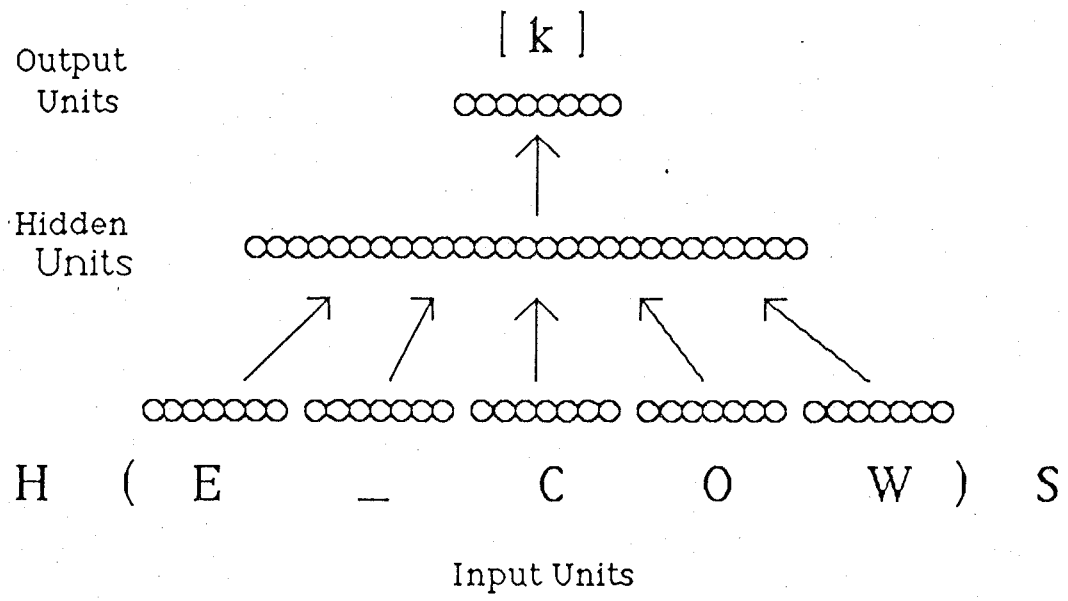

Figure 1. A schematic drawing of the NETtalk architecture. The little circles represent units (there are many more units than shown here) and the arrows represent bundles of connections, or weights, between the groups of units. Connectivity is complete between the connected groups, so each unit in each of the five input groups shown has connections to all of the units in the hidden layer, all of which are in turn connected to each of the output units. For the present experiments, there were 29 units in each of the input groups, 60 units in the hidden layer, and 26 units in the output layer. In addition, all units have a connection to a special unit that is always "on" (not shown), which serves as a variable threshold.

connections, which are real-valued and directional, determine how the activity of one unit affects the activity of another unit. If unit A is in an active state and there is a positively-valued (excitatory) connection going from unit A to unit B, then the activity level of unit B will be driven towards one. Conversely, a negatively-valued (inhibitory) connection between the two units will drive the activity level in unit B towards zero. A given unit typically has connections to a large number of other such units. The behavior of the network is the collective result of a large number of these simple, local, computations that are performed in parallel.

NETtalk has access to the correct pronunciations of the words during the learning, so it is "supervised" and akin to learning with a teacher. The back-propagation of error was used to adjust the values of the connection strengths (Rumelhart, Hinton, & Williams, 1986), which is a generalization of the perceptron learning rule (Rosenblatt, 1962) to multi-layered networks.

There are 231 units and 10,346 connections in the version of NETtalk used in the present experiments. As shown in Figure 1, the units that compose NETtalk are arranged in a layered hierarchy, consisting of three layers: an input layer, which encodes letters, an output layer, which encodes phonemes and stress, and a hidden layer that connects the input layer to the output layer. Each of the layers is completely connected to the layer just above and/or just below it. Letters are "clamped" at the input layer, and information (in the form of unit activity levels) passes up through the hidden layer, finally reaching the output layer where the pattern of activity on the output units is interpreted as a phoneme and stress.

The decision of how each letter is to be pronounced must be made on the basis of the surrounding letter context, since all letters can be pronounced in several ways. Using NETtalk, we have been able to examine how performance varies with window size. In this version, the network "sees" five letters at a time: the current letter, the preceding two letters, and the following two letters.[4] Each of these five letters is encoded simultaneously in a set or group of twenty-six dedicated units, locally representing each of the twenty-six English letters. Imposed on the network is a control structure that steps this five-letter window through the corpus, letter-by-letter.

More specifically, the value of each unit is a function of the values of all the units in the layer below it and the strength of the connection between the two units. The value of the ith unit is determined by first summing all of its inputs

$$E_i = \sum_j w_{ij} p_j \tag{1}$$

where p_j is the value of the jth unit and w_{ij} is the weight value of the connection between the two units, and then applying a sigmoidal transformation

$$p_i = P(E_i) = \frac{1}{1 + e^{-E_i}} \tag{2}$$

The resulting pattern of activity produced at the output layer is interpreted as the "guess" of how the middle letter in the window should be pronounced.[5] This output vector is then compared with the "correct" phoneme provided it, and the connection strengths in the network

[4] This window size has been reduced from seven in the original NETtalk in order to speed training.

[5] This was done by computing the projection of the output vector on all the possible phonemes (there are 55 of them) and selecting the phoneme with the highest overlap.

are recursively adjusted to minimize their differences (see Rumelhart, Hinton, & Williams, 1986, for details).

There were two adjustable learning parameters in the model: The rate of learning, ε, was set to 4.0, and the smoothing parameter, α, was set to zero (see Sejnowski & Rosenberg, 1986, for an explanation of these parameters). Continuous decay of the weight values towards zero has been experimented with, but was not used in the present experiments.

The purpose of the present experiment was to investigate the spacing effect in NETtalk, a network with two layers of modifiable weights. The design was modeled after Experiment 1 by Glenberg (1976). In this experiment, subjects were presented with paired associates, repeated twice at spacings of approximately 0, 1, 4, 8, 20, and 40 intervening items, and tested at retention intervals of approximately 2, 8, 32, and 64 items. Each pair was composed of two four-letter common nouns, "constructed to avoid common pre-experimental associations, rhymes, and orthographic similarities" (pg. 4). At test, just the stimulus word was presented, and the subject was to recall the associated response term. Glenberg's results are reproduced here as Figure 2. A significant interaction was found between spacing (lag) and retention interval. At short retention intervals, massed repetitions led to a higher probability of recall, whereas at long retention intervals, distributed repetitions were advantageous. Glenberg also noted that retention at the 64-item retention interval was a monotonic and negatively accelerating function of spacing.

As in Glenberg's experiment, the retention of target stimuli repeated a certain number of times at various spacing intervals was measured as a function of retention interval. If NETtalk exhibits the spacing effect, then long-term retention of these items should be better when a large number of other items intervene between successive repeats of the target (distributed practice). Conversely, short-term retention of the target items should be better when fewer items are presented between repeats (massed practice).

METHOD

Pre-Experimental Training

The network was first trained to pronounce a set of commonly occurring English words. These words were obtained by selecting the one thousand most frequent words from the *Webster's Pocket Dictionary,* based on frequency counts in the Brown corpus (Kucera & Francis, 1967). The network cycled through this one thousand word corpus a total of eleven times. The performance of the network at this point in training, as determined by the percentage of the correct phonemes "guessed", was 85%, and could have been improved with

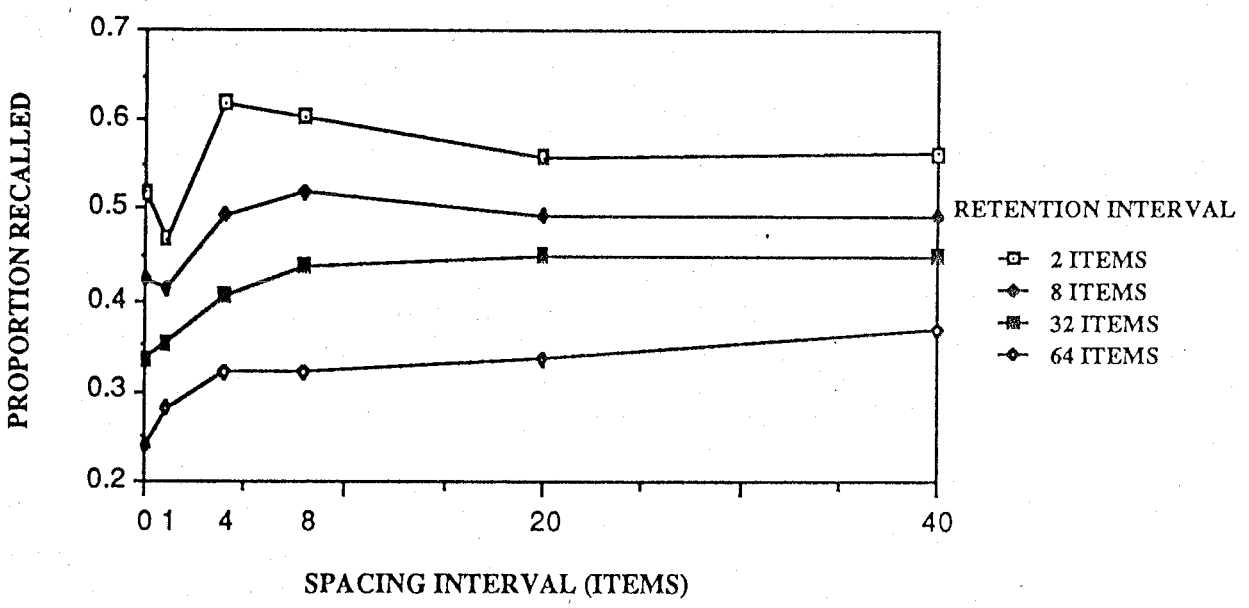

Figure 2. The proportion of response terms recalled as a function of spacing interval and retention interval. (After Glenberg, 1976.)

further practice.

The weight values of the network were stored following this initial training, and served as a common starting point for all of the subsequent experimental trials.

Target Stimuli

In order to force new learning to take place, random character strings of length six were employed as target stimuli. Thus there was no orderly relation between the cue and response. Whatever performance level NETtalk was able to reach on these items could not have been due to the utilization of rules acquired either prior or subsequent to study.

Twenty six-letter cues were generated by choosing six letters at random (with replacement) out of the twenty-six letters of the English alphabet. Likewise, the response terms associated with each of these cues were randomly generated phoneme and stress strings, also six characters each in length. There were 53 possible phonemes and five possible stress characters.[6] The frequency of occurrence of the characters in natural language were not taken into account in this selection process. Some of these items and several items from the

[6] In generating the target stimuli, two "phonemes", the space between words (_) and the period (.), were not possible choices.

Table 1. Examples of some training (distractor) and target items used.

DISTRACTORS

letters	phonemes	stress
file	fAl-	>1<<
all	cl-	1<<
second	sEkxnd	>1<0<<
take	tek-	>1<<
together	txgED--R	>0>1<<0<
neck	nEk-	>1<<
atmosphere	@tmxsf-Ir-	1<>0>>>2<<

RANDOM TARGET ITEMS

letters	phonemes	stress
fozepd	WdicnK	1<121>
sccfyk	p-UdSp	>202<1
bmyqcl	bzgTlz	0>><<>
grtufh	KCczOL	>1<010
eqhxxu	ANT|vM	>01<>2
ncssvr	zTSdWg	<<12>2
wxsale	RKpfI|	1<1110
djzxde	Yby^yI	20>>2>
kmfjqi	WGenGN	1><102

training corpus are presented in Table 1.

Procedure

The twenty target items were tested individually on separate trials. A trial consisted of first reading in the pre-experimental weights (described above), presenting a target item either two, ten, or twenty times, and then measuring the retention of the target as it was interfered with by subsequent learning. Furthermore, each target was presented at each of six spacing intervals, with either 0 (massed), 1, 4, 8, 20, or 40 (distributed) intervening items. Thus, eighteen trials were devoted to each target item (3 repetition groups x 6 spacing intervals). Between successive repeats of the target, words were presented from the original training corpus. Following the last repeat, the training corpus was again presented, and retention of the response terms of the target item was assessed after every item by presenting the cue term and measuring the mean squared difference between the output of the network and the correct

response

$$error = \frac{\sum_{j=1}^{J} (p_j^* - p_j)^2}{J} \tag{3}$$

for the J units in the output layer, where p_j^* is the target activation of the jth output unit, and p_j is it's actual value. Response accuracy was defined as one minus the mean error for the word

$$accuracy = 1 - \frac{\sum_{n=1}^{N} error}{N} \tag{4}$$

where N is the number of letters in the word. Learning was turned off (achieved by setting the learning rate to zero) for these tests, so that no changes were made to the strengths of the connections in the network.

RESULTS

Accuracy, as defined above, was averaged over the twenty target items and plotted as a function of retention interval for each repetition group (Figure 3). Following Glenberg (1976), values were selected from this curve at retention intervals of 2, 8, 32, and 64 items and re-plotted as a function of spacing interval (Figure 4).

A 6 (spacing intervals) x 4 (retention intervals) analysis of variance was performed on these selected values, treating target items as subjects. The main effect of retention interval was highly significant in all repetition groups, F (3, 57) = 32.82, 58.50, and 48.29, all $p < 0.001$, for the two, ten, and twenty repetition groups, respectively, indicating that a considerable amount of forgetting of the target items did take place. The main effect of spacing was highly significant only in the twenty repetition condition, F (5, 95) = 5.10, $p < 0.001$, and marginally significant in the ten repetition condition, F (5, 95) = 3.02, $p < 0.05$. Of interest, however, was the interaction between spacing and retention interval. This interaction was significant for all three repetition groups: F (15, 285) = 27.68 and 37.29, both $p < 0.001$, for the ten and twenty repetition conditions, respectively, and F (15, 285) = 2.73, $p < 0.03$, in the two repetition condition.

A trends analysis of the accuracy measures was performed across spacings for retention intervals of 0 (short-term) and 64 (long-term) items. The downward trend in retention for immediate retention as spacing increased was highly significant following ten and twenty repeats of the target, F (5, 95) = 17.14 for the ten, and F (5, 95) = 6.70 for the twenty repetition groups, both $p < 0.001$. In both cases, the linear trend was highly significant, F (1,

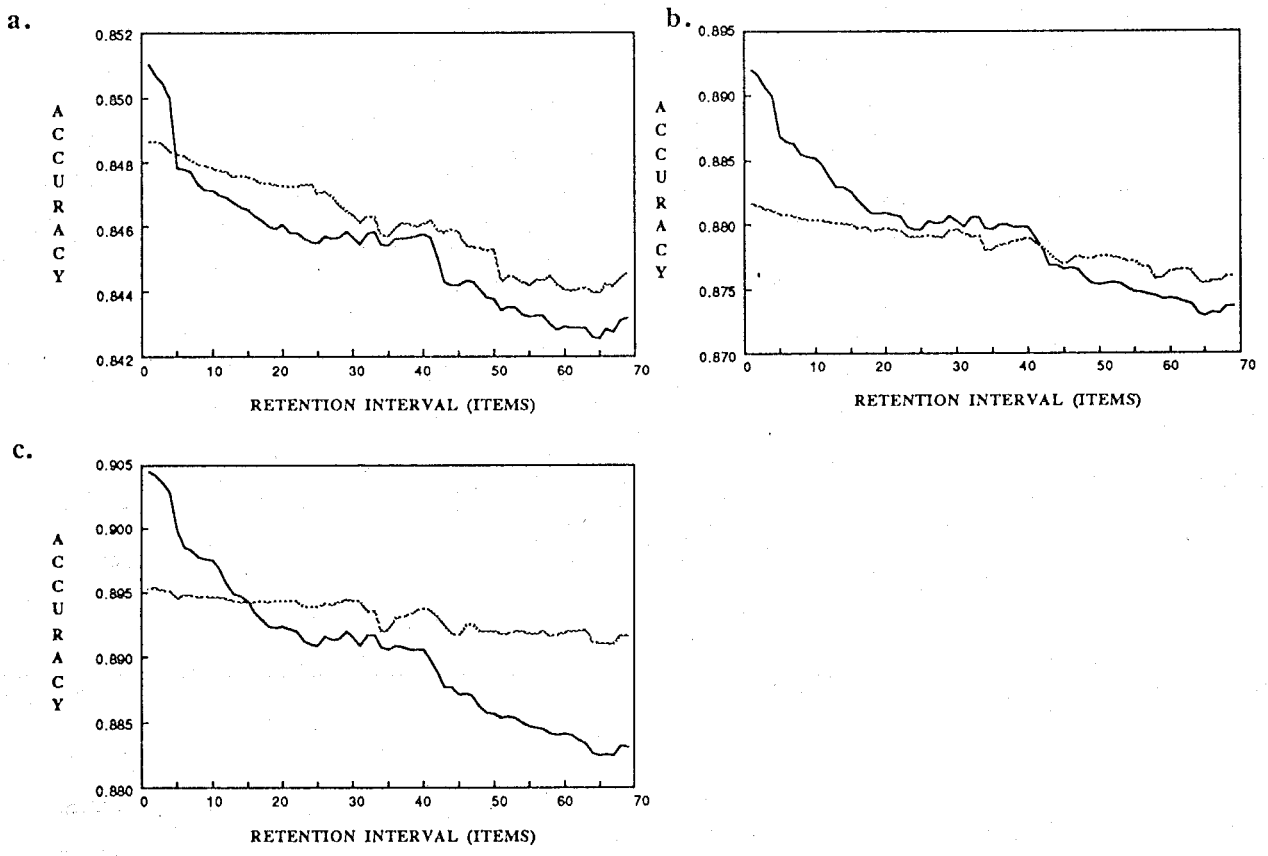

Figure 3. Mean response accuracy over all target items plotted as a function of retention interval following two (a), ten (b), and twenty (c) repetitions of the target item. Only spacing intervals of zero (solid) and forty (dots) items are shown.

19) = 38.43, $p < 0.001$, and F (1, 19) = 11.80, $p < 0.001$, for the ten and twenty repetition groups, respectively. This downward trend was not significant after only two repeats, however, F (5,15) < 0.5. Neither the quadratic nor the cubic trends reached significant levels in any of the three repetition groups.

The upward trend at the 64-item retention interval was significant for all repetition groups, F (5, 95) = 2.31, $p = 0.05$, for two repeats, and F (5, 95) = 8.92 and 22.40, both $p < 0.001$ for the ten and twenty repetition groups. The shape of the curve varied, however. As the number of repetitions increased from two to ten to twenty, the trend varied from cubic, F (1, 19), $p < 0.05$, to linear, quadratic, and cubic, F (1, 19) = 6.06, 24.94, and 6.12, $p < 0.05$, 0.001, and 0.05, respectively, to only linear and quadratic, F (1, 19) = 16.21 and 66.63, both $p < 0.001$.

DISCUSSION

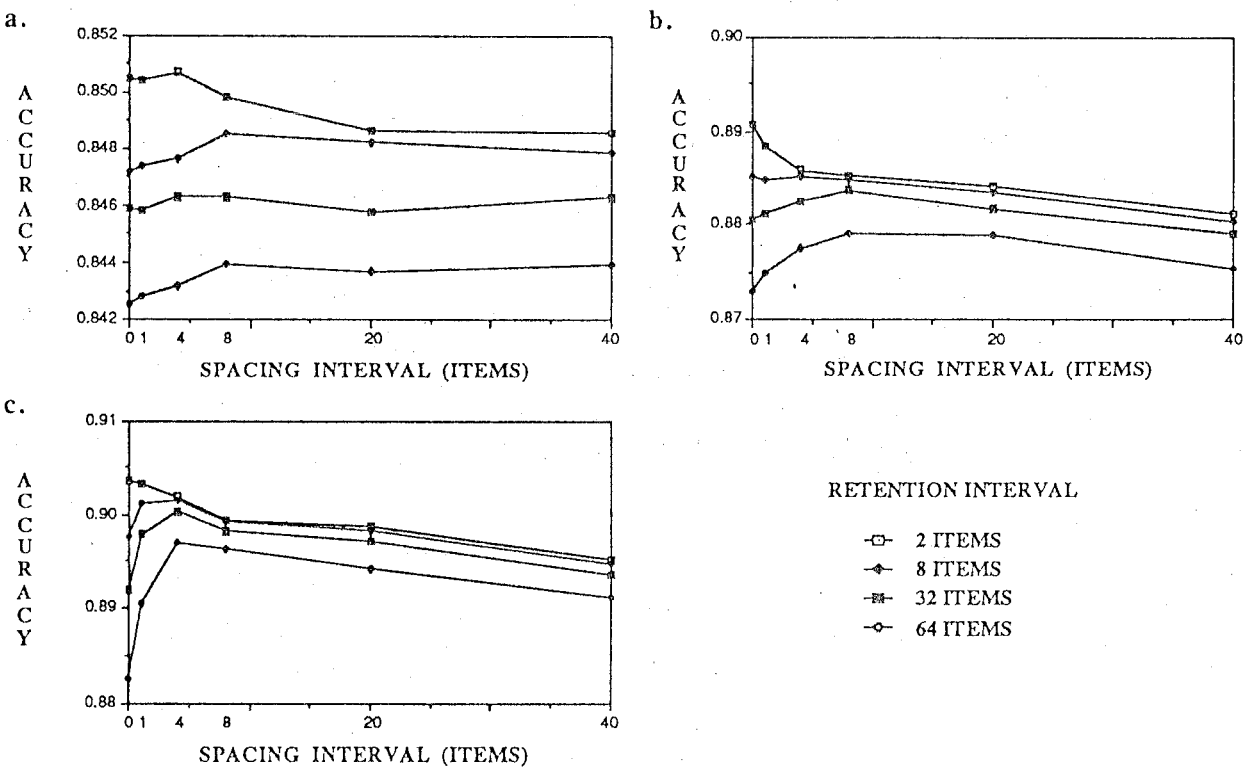

Figure 4. Mean response accuracy plotted as a function of spacing interval at 2, 8, 32, and 64 item retention intervals for the two (a), ten (b) and twenty (c) repetition groups.

A significant spacing effect was observed in NETtalk: Retention of nonwords after a 64-item retention interval was significantly better when presented at the longer spacings (distributed presentation) than at the shorter spacings. In addition, a significant advantage for massed presentations was found for short-term retention of the items. Although stimulus materials, response measures, and procedure differ sufficiently to make direct comparison impossible, the overall pattern of these results resembles that found by Glenberg (1976), in an experiment using human subjects. We obtained our results without making additional assumptions or including additional mechanisms such as consolidation, rehearsal, or attention. Nor were explicit assumptions made about a continuously changing context other than the context implicitly provided by the network.

Recency effects, similar to those reported here, are common in the human literature and have been reported in spacing experiments (e.g. Peterson, Wampler, Kirkpatrick, & Saltzman, 1963; Sperber, 1974). This short-term advantage for massed practice is commonly discussed with reference to a limited-capacity memory buffer. The present experiments indicate that some of the effects such a mechanism was designed to account for can be produced without such a device.

Why should NETtalk exhibit these characteristics? The answer, as we attempt to show, depends on the way in which learning and the resulting knowledge is represented in NETtalk.

An intuitive understanding of learning in NETtalk can be obtained by thinking of the set of n weight values, the many sites of learning in the network, as specifying a point in an n-dimensional hyperspace. The goal of learning is to move on a trajectory through this hyperspace towards a point where the error on the entire training corpus is minimized. The direction in which to move at any one point is determined by estimating the error gradient. If a global minimum for a given corpus can be reached, no further learning (i.e. weight adjustments) is required. A minimum is consequently a point of high stability, *until* a new item is presented that is irregular, or is for some other reason not like the other items in the training corpus. Our hypothesis is that distributing practice leads to a more stable position in this hyperspace upon the re-presentation of the training corpus.

For the sake of simplicity, suppose that NETtalk has only three connections, so that its state at any one time can easily be represented in a 3-dimensional space (see Figure 5). Suppose further that, as in the present simulations, this network has been trained on a large pre-experimental training corpus and that it has reached an optimum (where the error is at a global minimum) for these items (Point A). Now a new and unusual target item is presented in either a massed or spaced condition to our mini-network. If the target is presented several times back-to-back, as in the massed condition, minimizing the error following each presentation will lead us down a path toward a point that is optimal for this target item, perhaps even reaching this optimum (Point B). But because this voyage will have taken us quite out of the way from our starting point (A), this new position is not likely to be stable to the re-presentation of the training corpus, and so the massed learning of the new item will be lost quickly.

Assuming, however, that there *is* a point that is optimal for both the training corpus and the target item (Point C in the figure), alternating presentations of the target with items from the training set is one way of moving closer to this highly stable point.[7] Upon the first presentation of the target item, the error gradient for that item is estimated and the error is reduced by adjusting the weights in the direction of the steepest descent (to position 1). So far, this procedure has been identical to that for the massed condition, and so the network is at

[7] Another way is to update the weight values less frequently. Instead of learning in small increments, as in NETtalk, which updates after every word, one could also collect data over many trials and then take one big jump. Although this procedure (within its resolution) overcomes the problems associated with presentation order (such as the spacing effect), it may be a hazardous one, since new information is integrated at a slow rate.

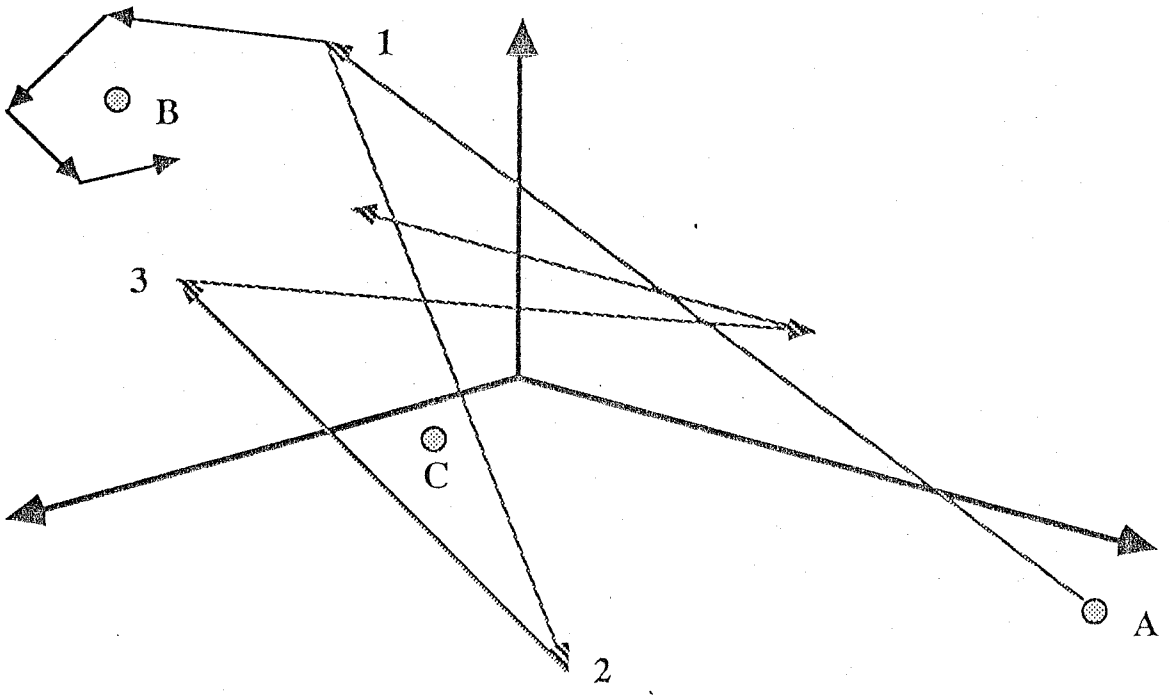

Figure 5. Movements in weight-space during learning for massed (solid) and distributed (textured) conditions. Point A is a global optimum for the pre-experimental training corpus (the assumed starting point for all experimental trials), Point B is an optimum for the target item, and Point C is an optimum for both the target and the training corpus. (See text for explanation.)

the same point in hyperspace. Now, however, instead of presenting the target again, an item from the original training corpus is presented. Again, the weights are adjusted to minimize the error on the item (to position 2), only this time the direction of movement is more likely to be towards Point A than Point B, since A was a global minimum for the training corpus. Presenting the target again will cause a movement back towards B (to position 3), and so on. We see that distributing practice causes the network to weave back and forth in this hyperspace, allowing it to perform a more complete search of the error space for *both* the training corpus *and* the target item. The network therefore has a better chance of finding the optimal position (Point C) than it would if practice were massed, and its encoding of the target item will consequently be more able to withstand interference due to further training on both types of material.

The explanation of the spacing effect that we offer here is not meant as an alternative to previous suggestions; it is a different type of explanation, relying as it does on the underlying structure of the representations. The decline in learning rate as local optima are approached is reminiscent of the process of habituation: less is effectively learned each time the item is

repeated. Other aspects of our model bear a resemblance to encoding variability to the extent that items are encoded relative to the current state of the network, which is in a state of continual flux. And if we identify Jacoby's processing effort with the degree of change required to construct a distributed representation, then our simulations can be considered support for this proposal as well. Nevertheless, while these concepts of habituation, encoding variability, and processing effort may be reinterpreted within the framework of connectionist models such as ours, they are at a different level of explanation.

Our results are limited to a particular network architecture in a particular domain. To what extent is this conclusion dependent on the details of our model? If the spacing effect is a direct consequence of incremental learning in memory systems that use distributed representations, as we suspect, then the same effects of massed and distributed learning should occur in other task domains and with other network architectures that also have learning algorithms with distributed representations, such as Boltzmann machines (Hinton & Sejnowski, 1983; Ackley, Hinton, & Sejnowski, 1985). We predict as well that the same general principles may underlie the spacing effect in human learning.

References

Ackley, D.H., Hinton, G.E. , & Sejnowski, T.J. (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science, 9*, 147-169.

Amari, S. (1977). Neural Theory of Association and Concept Formation. *Biological Cybernetics, 26*, 175-185.

Anderson, J.A., Silverstein, J.W., Ritz, S.A., & Jones, R.S. (1977). Distinctive Features, Categorical Perception, and Probability Learning: Some Applications of a Neural Model. *Psychological Review, 84*, 413-451.

Atkinson, R.C., & Shiffron, R.M. (1968). Human Memory: A Proposed System and Its Control Processes. In K.W. Spence & J.T. Spence (Eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 2). New York: Academic Press.

Ballard, D.H., Hinton, G.E. , & Sejnowski, T.J. (1983). Parallel visual computation. *Nature, 306*, 21-26.

Bjork, R.A., & Allen, T.W. (1970). The Spacing Effect: Consolidation or Differential Encoding. *Journal of Verbal Learning and Verbal Behavior, 9*, 567-572.

Ebbinghaus, H. (1964). *Memory: A Contribution to Experimental Psychology* (originally published, 1885) . New York: Dover.

Estes, W.K. (1959). The Statistical Approach to Learning Theory. In S. Koch (Ed.), *Psychology: A Study of a Science* (Vol. II). New York: McGraw-Hill.

Feldman, J.A., & Ballard, D. (1982). Connectionist Models and Their Properties. *Cognitive Science, 6*, 205-254.

Glenberg, A.M. (1976). Monotonic and Nonmonotonic Lag Effects in Paired-Associate and Recognition Memory Paradigms. *Journal of Verbal Learning and Verbal Behavior, 15*, 1-16.

Glenberg, A.M. (1979). Component-levels Theory of the Effects of Spacing of Repetitions on Recall and Recognition. *Memory and Cognition, 7*, 95-112.

Hinton, G.E., & Sejnowski, T.J. (1983). *Optimal Perceptual Inference.* Washington, D. C.: Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition.

Hintzman, D.L. (1974). Theoretical Implications of the Spacing Effect. In R.L. Solso (Ed.), *Theories in Cognitive Psychology: The Loyola Symposium.* Hillsdale, N.J.: Erlbaum.

Hintzman, D.L. (1976). Repetition and Memory. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 10). New York: Academic Press.

Hintzman, D.L., Block, R.A., & Summers, J.J. (1973). Modality Tags and Memory for Repetitions: Locus of the Spacing Effect. *Journal of Verbal Learning and Verbal Behavior, 12*, 229-238.

Hintzman, D.L., Summers, J.J., & Block, R.A. (1975). What Causes the Spacing Effect? Some Effects of Repetition, Duration, and Spacing on Memory for Pictures. *Memory and Cognition, 3*, 287-294.

Hintzman, D.L., Summers, J.J., Eki, N.T., & Moore, M.D. (1975). Voluntary Attention and the Spacing Effect. *Memory and Cognition, 3*, 576-580.

Jacoby, L.L. (1978). On Interpreting the Effects of Repetition: Solving a Problem Versus Remembering a Solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649-667.

Kohonen, T., Oja, E., & Lehtio, P. (1981). Storage and Processing of Information in Distributed Associative Memory Systems. In G.E. Hinton & J.A. Anderson (Eds.), *Parallel Models of Associative Memory.* Hillsdale, N.J.: Erlbaum.

Kucera, H., & Francis, W.N. (1967). *Computational Analysis of Modern-Day American English.* Providence, R.I.: Brown University Press.

Landauer, T.K. (1969). Reinforcement as Consolidation. *Psychological Review, 76*, 82-96.

Maki, R.H., & Hasher, L. (1975). Encoding Variability: A Role in Immediate and Long-term Memory?. *American Journal of Psychology, 88*, 217-231.

Martin, E. (1968). Stimulus Meaningfulness and Paired-Associate Transfer: An Encoding Variability Hypothesis. *Psychological Review, 75*, 421-441.

McClelland, J.L., & Rumelhart, D.E. (1985). Distributed Memory and the Representation of General and Specific Information. *Journal of Experimental Psychology: General, 114*, 159-188.

Melton, A.W. (1970). The Situation with Respect to the Spacing of Repetitions and Memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 596-606.

Peterson, L.R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of Spacing Presentations on Retention of a Paired-Associate Over Short Intervals. *Journal of Experimental Psychology, 66*, 206-209.

Postman, L., & Knecht, K. (1983). Encoding Variability and Retention. *Journal of Verbal Learning and Verbal Behavior, 22*, 133-152.

Rosenblatt, F. (1962). *Principles of Neurodynamics.* Washington, D.C.: Spartan Books.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning Internal Representations by Error Propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, Mass.: MIT

Press.

Rundus, D. (1971). Analysis of Rehearsal Processes in Free Recall. *Journal of Experimental Psychology, 89*, 63-77.

Sejnowski, T.J., & Rosenberg, C.R., NETtalk: A Parallel Network that Learns to Read Aloud, The Johns Hopkins University Electrical Engineering and Computer Science Technical Report JHU/EECS-86/01, 1986.

Shaughnessy, J.J., Zimmerman, J., & Underwood, B.J. (1972). Further Evidence on the MP-DP Effect in Free-Recall Learning. *Journal of Verbal Learning and Verbal Behavior, 11*, 1-12.

Sperber, R.D. (1974). Developmental Changes in Effects of Spacing of Trials in Retardate Discrimination Learning and Memory. *Journal of Experimental Psychology, 103*, 204-210.

Tzeng, O.J.L. (1973). Stimulus Meaningfulness, Encoding Variability, and the Spacing Effect. *Journal of Experimental Psychology, 99*, 162-166.