Journal of Experimental Psychology:
Learning, Memory, and Cognition
2000, Vol. 26, No. 5, 1103-1123

Copyright 2000 by the American Psychological Association, Inc.
0278-7393/00/$5.00 DOI: 10.1037//0278-7393.26.5.1103

Age of Acquisition Effects in Adult Lexical Processing Reflect Loss of Plasticity in Maturing Systems: Insights From Connectionist Networks

Andrew W. Ellis
University of York

Matthew A. Lambon Ralph
Medical Research Council Cognition and Brain Sciences Unit

Early learned words are recognized and produced faster than later learned words. The authors showed that such age of acquisition effects are a natural property of connectionist models trained by back-propagation when patterns are introduced at different points into training and learning of early and late patterns is cumulative and interleaved. Analysis of hidden unit activations indicated that the age of acquisition effect reflects a gradual reduction in network plasticity and a consequent failure to differentiate late items as effectively as early ones. Further simulations examined the effects of vocabulary size, learning rate, sparseness of coding, use of a modified learning algorithm, loss of early items, acquisition of very late items, and lesioning the network. The relationship between age of acquisition and word frequency was explored, including analyses of how the relative influence of these factors is modulated by introducing weight decay.

All other things being equal, words learned early in life are recognized and produced faster and more accurately than words learned later. This is true for a variety of lexical processing tasks, including object naming, word naming, visual lexical decision, and auditory lexical decision (Barry, Morrison, & Ellis, 1997; Carroll & White, 1973; Ellis & Morrison, 1998; Gerhand & Barry, 1998, 1999a, 1999b; Gilhooly & Gilhooly, 1979; Morrison & Ellis, 1995, 2000; Turner, Valentine, & Ellis, 1998).

Carroll and White (1973) were the first to relate word learning age to adult processing speed. They showed that object naming latency correlated .77 with a measure of the age at which children learn the different object names. In a multiple regression analysis, age of acquisition was the only significant independent predictor of naming latency. Their demonstration of an effect of age of acquisition on object naming speed has been replicated several times since (e.g., Barry et al., 1997; Cuetos, Ellis, & Alvarez, 1999; Gilhooly & Gilhooly, 1979; Morrison, Ellis, & Quinlan, 1992; Snodgrass & Yuditsky, 1996; Vitkovitch & Tyrell, 1995) and has been shown to hold when objective measures of the age at which different words are learned are used in place of subjective (rated) measures (Ellis & Morrison, 1998; Morrison, Chappell, & Ellis, 1997). In all of these studies, age of acquisition showed the highest correlation with naming latency of any variable investigated. Age of acquisition has also been shown to predict which objects will prove easy or hard to name for brain-injured patients with aphasia or semantic dementia (Ellis, Lum, & Lambon Ralph, 1996; Hirsh & Ellis, 1994; Hirsh & Funnell, 1995; Lambon Ralph, Graham, Ellis, & Hodges, 1998; Nickels & Howard, 1995), as well as for elderly people experiencing normal, age-related problems with word finding (Hodgson & Ellis, 1998). In each case, later learned words were found to be more vulnerable to damage and decay than early learned words.

Despite the abundant evidence that age of acquisition is an important determinant of lexical processing speed, Ellis and Morrison (1998) noted a paucity of theoretical proposals for precisely how age of acquisition might come to influence adult lexical processing. Gilhooly and Watson (1981) argued that if object naming latency is affected by age of acquisition rather than word frequency, then age of acquisition rather than frequency may be held to determine the thresholds of lexical units involved in word recognition and production (the logogens of Morton's, 1969, model). Brown and Watson (1987) proposed that the phonological representations of early acquired words may be stored in unitary form but that the phonological representations of later acquired words may be more fragmentary in nature. The extra processing time required to assemble the phonological form of a late acquired word could account for the slower processing of such words. This proposal is still cited regularly (e.g., Barry et al., 1997; Gerhand & Barry, 1999b; Morrison & Ellis, 1995) but has never been subjected to direct empirical test.

A form of explanation that has, until now, been lacking for age of acquisition is one couched in terms of the performance of connectionist networks. This situation contrasts starkly with the ease with which, for example, frequency effects in word recognition and production have been incorporated into connectionist theories. Morton's (1969) notion that frequency affected the thresholds of word recognition units (logogens) could be readily

Andrew W. Ellis, Department of Psychology, University of York, York, England; Matthew A. Lambon Ralph, Memory Group, Medical Research Council Cognition and Brain Sciences Unit, Cambridge, England.

The research reported in this article was supported by Grant G9305476N from the Medical Research Council of Great Britain and by a grant from the National Institutes of Mental Health to K. Patterson and J. L. McClelland. We thank Mike Masson and an anonymous referee for helpful suggestions and J. L. McClelland for his invaluable assistance with the analysis of network performance.

Correspondence concerning this article should be addressed to Andrew W. Ellis, Department of Psychology, University of York, York YO10 5DD, England. Electronic mail may be sent to awe1@york.ac.uk.

translated into the design of early connectionist models using localist representations (e.g., McClelland & Rumelhart, 1981). Later models, including those that used distributed rather than localist representations, generally moved to the position that frequency is embodied in the strength of connections between representations, so each encounter with a word strengthens the links between the different representations (e.g., semantic, phonological, and orthographic) involved in recognizing and producing the word (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). Spelling–sound regularity and word imageability are two other factors whose influence on lexical processing in normal or brain-injured individuals, or both, has been given a plausible account in terms of connectionist models (e.g., Plaut & Shallice, 1993; Plaut et al., 1996).

In contrast, age of acquisition effects have been held by some authors to be positively at odds with current connectionist models (e.g., Gerhand & Barry, 1998; Moore & Valentine, 1998; Morrison & Ellis, 1995). This belief has been based on comparisons (which we now consider to be false) between age of acquisition effects in humans and the phenomenon of "catastrophic interference" in neural networks (Lewandowsky, 1991; McCloskey & Cohen, 1989; Ratcliff, 1990; Sharkey & Sharkey, 1995). If neural networks are trained on one set of patterns, which are then replaced in training by a different set of patterns, performance on the first set suffers. Knowledge of the first set, which is no longer being trained, is gradually lost as the second set is learned. If the first training set is held to be analogous to early learned words and the second set to later learned words, then catastrophic interference produces exactly the opposite results from human age of acquisition: The later patterns displace the early ones, whose representations deteriorate steadily.

But when a child acquires a vocabulary under natural conditions, new words do not supplant and replace preexisting words; rather, the child's vocabulary grows cumulatively, with new (later acquired) words being added gradually to the existing stock of old (earlier acquired) words. Those early acquired words do not cease to be used or encountered but are interspersed among, and interleaved with, the later acquired words. McClelland, McNaughton, and O'Reilly (1995) noted the importance of gradual interleaved learning for effective acquisition of new material by networks trained by back-propagation. Using a network devised by Rumelhart and Todd (1993) that was trained to reproduce the correct semantic propositions for a small selection of concepts (trees, flowers, birds, and fish), McClelland et al. (1995) examined the consequences of introducing a novel concept (penguin) after the network had been trained. They did this using either *focused* or *interleaved* learning. With focused learning, new knowledge is presented to the system without interleaving it with old knowledge and without continued exposure to the earlier training set. Under such conditions, information about penguins was acquired rapidly but at a cost to preexisting knowledge. In other words, the model showed catastrophic interference. If, however, learning was interleaved (or, as stated here, *cumulative*), with the model being given continued exposure to the old material alongside the new, then the new information was acquired without cost to the old.

Catastrophic interference may therefore be avoided in circumstances in which learning is cumulative, as is the case with vocabulary acquisition in childhood. For a neural network to provide a viable and intuitively plausible account of age of acquisition effects in humans, one would need to train the network on one set of patterns (early acquired) and then introduce a second set (late acquired) under conditions of cumulative, interleaved training. It would be necessary to demonstrate that catastrophic interference does not occur and that, even after extensive training, the network processes the early items more efficiently than the later ones. Those are the aims of this article. We focus our attention on networks that use distributed representations and in which learning is accomplished by the application of the back-propagation algorithm, although we argue later that effects of the sort reported here may prove to be ubiquitous. Distributed memory networks trained by back-propagation have been widely used to simulate the recognition, understanding, and generation of words. In contrast to the claims of Gerhand and Barry (1998), Moore and Valentine (1998), and Morrison and Ellis (1995), we show that age of acquisition effects can readily be simulated in such networks. In doing so, we also show that age of acquisition effects cannot be explained in terms of differences between early and late pattern sets in their cumulative frequency of training. We then report analyses based on hidden unit activations that provide clues as to the origins of differences between early and late patterns after training on both. Further simulations explored the effects of varying sparseness of coding and vocabulary size in the model and the consequences for the behavior of the model of changing the learning rate and training with a modified learning algorithm (Quickprop). We then examine the relationship between the point of entry of pattern sets into training (age of acquisition) and the frequency with which those patterns are trained. These simulations included analyses of what happens when early items are lost from training and when the mature model is required to learn very late sets of patterns. The final simulations examined the effects of introducing weight decay and of damaging the fully trained network. On the basis of these simulations, we argue that not only do age of acquisition effects arise as a natural consequence of training with back-propagation, but such effects may be an inescapable feature of neural networks that learn by changing weights on connections in such a way that early training has a greater impact on network structure than later training.

Simulations 1 and 2: Consequences for Network Performance of Entering Patterns Into Training at Different Points Under Either Cumulative (Interleaved) or Focused Conditions

Our first set of simulations investigated the effects of introducing sets of patterns at different points in the training of a standard three-layer back-propagation network under conditions of either cumulative, interleaved training, in which the late set was added into training after a period of training on the early set only (Simulation 1), or focused learning in which the late set replaced the early set in training (Simulation 2). Input and output layers of the network each contained 100 units that were fully connected to an intermediate, hidden layer containing 50 units. The model was trained on 200 different patterns (100 early and 100 late). The input representations were randomly generated, binary patterns distributed across all 100 input units. Most bits within each pattern were set to 0, but some were given a value of 1 (with a probability of .2). Each output pattern was a copy of the corresponding input pattern after a degree of perturbation had been applied (bit values

were flipped from 0 to 1, or vice versa, with a probability of .1). Input and output patterns were therefore correlated. The model was trained to produce each of the 200 correct output patterns from the corresponding 200 input patterns. Weights on each connection of the model were initialized to small random values and were adjusted in the direction of a maximum of 1 or a minimum of −1 after the presentation of each pattern with the back-propagation learning algorithm (using the *bp* software; McClelland & Rumelhart, 1988). Both models were trained with a fixed learning rate (0.05) and no momentum (i.e., weight change was computed for the current pattern only without reference to prior weight adjustments). Thus, the parameters that relate to the mechanisms of learning remained constant from the first epoch of training to the last. Unless otherwise stated, all of the simulations reported here used this network and patterns with these characteristics.

The 200 input–output patterns were divided into two sets, 100 early and 100 late. Training was given in two forms. In Simulation 1 (cumulative, interleaved learning), the 100 early patterns were first trained for 250 epochs (where 1 epoch involves a single presentation of each pair of input–output patterns). The 100 late patterns were then added to the early patterns in the training set, and the model was given another 250 epochs of training on the combined corpus of 200 patterns. This measure reflects the difference between the network's output and the ideal response. The closer the value is to zero, the better the performance of the model.

Learning under cumulative training conditions was compared with two versions of focused learning. In Simulation 2.1, the 100 early patterns were again presented for 250 training epochs but were then wholly replaced by the 100 late patterns, which were trained for a further 250 epochs. Performance of the network was assessed after 500 epochs, by which point each pattern had been presented to the network 250 times. In Simulation 1, however, the early patterns had been presented 500 times by Epoch 500, in comparison with 250 times for the late patterns. A second version of focused learning was therefore run (Simulation 2.2) in which the early patterns were trained for 500 epochs before being wholly replaced by the late patterns for a further 250 epochs. The performance of the network was assessed at 750 epochs, by which time the total number of presentations of the early and late patterns was the same as in Simulation 1.

Figure 1 shows the performance of the network after 250 and 500 epochs for Simulation 1 (Figure 1A) and Simulation 2.1 (Figure 1B) and after 500 and 750 epochs for Simulation 2.2 (Figure 1C). When the late patterns were added to the early patterns in a cumulative manner (Simulation 1), the late patterns were incorporated into the training schedule, and at 500 epochs the model performed reasonably well on both sets of patterns, although better on the early than the late. This is similar to the network behavior described by McClelland et al. (1995). In contrast, both versions of focused learning caused the model to demonstrate loss of the early learned associations. This was true whether the early patterns were presented for 250 epochs (Figure 1B) or 500 epochs (Figure 1C) before being replaced by the late patterns. Had training on the late sets continued in Simulations 2.1 and 2.2 with no further presentations of the early set, knowledge of the early set would have become progressively weaker; that is, the loss of early patterns due to interference from the late patterns would have become catastrophic.

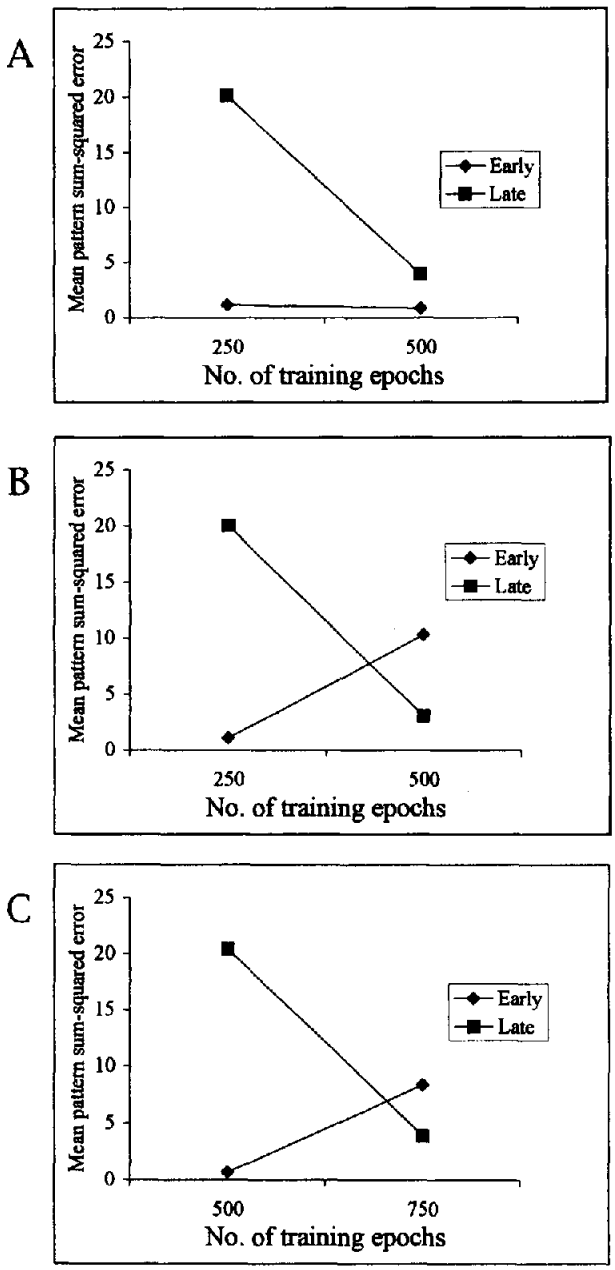

Figure 1. Effects of order of entry into a distributed network trained by back-propagation under conditions of cumulative, interleaved learning with late patterns being introduced after 250 epochs (A; Simulation 1) or under conditions of focused learning with late patterns being introduced after either 250 epochs (B; Simulation 2.1) or 500 epochs (C; Simulation 2.2).

Simulation 3: Extended Training on Early and Late Sets Under Conditions of Cumulative, Interleaved Training

In the cumulative learning situation of Simulation 1, performance after 500 epochs was worse on the late patterns introduced into training after 250 epochs than on the early patterns that had been trained from the outset. It is possible, however, that learning

had not yet approached asymptote by 500 epochs and that the performance on the early and late patterns would converge given sufficient additional training on both. We would argue that if this were to happen, it would not provide as plausible an account of age of acquisition effects in human lexical processing as would a situation in which the difference between early and late patterns persisted after extensive further training on both sets. This is because differences of a few years in the age of acquisition of words in childhood have consequences that can still be detected decades later (Ellis et al., 1996; Hodgson & Ellis, 1998; Lambon Ralph et al., 1998).

Simulation 3 extended Simulation 1 beyond 500 epochs up to 100,000 epochs. As in Simulation 1, the first 250 epochs involved training on the early set of 100 patterns only. From 250 to 100,000 epochs, the training corpus contained all 200 patterns. The results are shown in Figure 2. The performance of the network at 250 and 500 epochs was the same as in Figure 1A. The network continued to learn after 500 epochs, but reductions in error with additional training declined progressively. Thus, the reduction in error between 50,000 and 60,000 epochs was only 0.03%, and error declined by just 0.01% between 90,000 and 100,000 epochs. Even though error reduction was vanishingly small by 100,000 epochs, the consequences of a head start of just 250 epochs for the early set remained substantial; the mean sum-squared error was 0.23 for the early patterns, as compared with 2.48 for the late patterns, $t(98) = 11.5$. (Unless otherwise stated, we assume a significance level of .001 throughout.) The reduced error for early patterns trained from the outset relative to patterns introduced later constitutes, we believe, a plausible demonstration of an enduring age of acquisition effect in a distributed connectionist network of a type used widely to model adult word recognition and production.

Simulation 4: Age of Acquisition Does Not Reduce to Simple Differences in the Cumulative Frequency of Early and Late Patterns

Simulation 3 showed that the age of acquisition effect is not a simple matter of differences in the total frequencies with which the early and late sets are presented to the network. After 100,000 epochs of training, the cumulative frequencies for the two sets of patterns were almost identical (ratio of early to late = 1.003:1), yet there was a tenfold difference in error scores. The fact of having been entered early into training imparts an advantage to a set of patterns that remains more or less constant as long as the relative frequencies with which the different sets of words are encountered do not change (see Simulations 11–13 and 15). This implies that human age of acquisition effects will not reduce in any simple sense to cumulative frequency or "residence time" in lexical memory (Gilhooly, 1984; Lewis, 1999) and that once age of acquisition differences are established in childhood, they are likely to remain largely unchanged across the adult life span.

Simulation 4 constituted a direct test of the claim that two sets of patterns may be presented an equal number of times to the network, but if one set is introduced earlier than the other and continues to be presented after the second set has been introduced, then the early set will retain a processing advantage over the late set. In Simulation 4, the model was first trained on the 100 early patterns for 1,000 epochs, after which the 100 late patterns were introduced and trained alongside the early patterns for a further 1,000 epochs. The late patterns were, however, presented twice per epoch between 1,000 and 2,000 epochs while the presentation rate of the early patterns remained at once per epoch. After 2,000 epochs, the network had therefore been presented with each pattern, early or late, 2,000 times. Despite equating of the cumulative frequency of the early and late patterns, the network performed significantly better on the early items, with the mean pattern sum-squared error for the early patterns being 0.4, as compared with 3.3 for the late patterns, $t(198) = 14.8$. Thus, the same amount of training on an input–output pattern has different effects depending on when that training occurs.

Simulation 5: Staggering the Point of Entry of Pattern Sets Into Training

Human vocabulary acquisition does not involve learning one set of words for a period of time and then suddenly adding a second

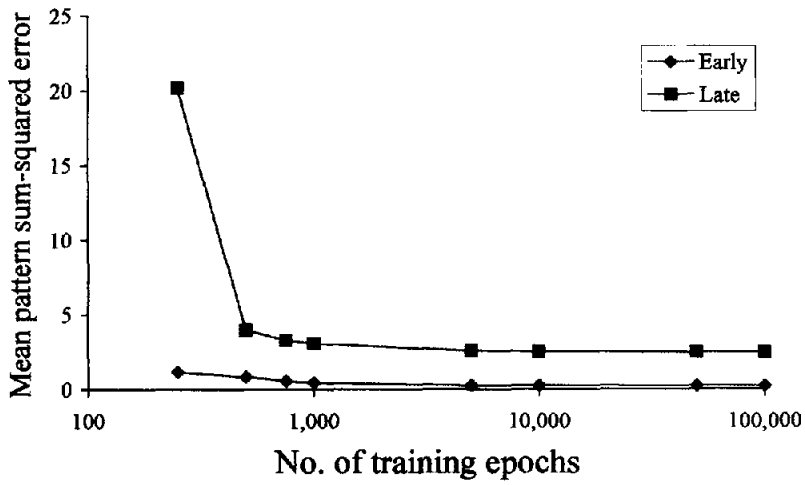

Figure 2. Effects of continued training on early and late patterns under conditions of cumulative, interleaved training (Simulation 3).

set; rather, natural vocabulary acquisition involves the gradual accumulation of words with constant interleaving of old with new items. This article is more concerned with illustrating the principle that connectionist networks can account for age of acquisition effects than with producing precise and detailed simulations of natural vocabulary acquisition. Simulation 5 did, however, examine the effects of adding patterns into training more gradually, with one set of very early patterns being followed by three further sets (early, medium, and late) entered one after the other under conditions of cumulative training.

With the same network as before, the 200 input–output patterns were divided into four sets of 50. The very early set was trained from the outset. The early set was added into training after 200 epochs, whereas the medium and late sets were added into training after 400 and 600 epochs, respectively. Figure 3 shows the performance of the network up to 5,000 epochs, by which time it had virtually stabilized. The consequence of each additional delay of 200 epochs was an average increase in mean pattern sum-squared error for that set of 1.16 at 5,000 epochs. A one-way analysis of variance carried out on the error scores at 5,000 epochs confirmed that the difference in error between sets was significant, $F(3, 196) = 68.5$, $MSE = 113.6$. Tukey's honestly significant difference post hoc comparisons between each set and the next were also significant. Thus, adding patterns gradually to the network results in a steady worsening of final performance on later relative to earlier sets.

How Does Age of Acquisition Affect Network Performance?

We have established that patterns on which the network is trained from the outset develop representations that generate better performance in the mature network than patterns entered later. This is true even when the difference in the delay between entering the early and the late patterns is small relative to the period for which the network is subsequently trained on both sets. The present section is concerned with trying to understand why it is that a small head start given to one set of patterns has such lasting consequences for the behavior of the network and makes it very difficult for patterns entered later to develop representations of equivalent quality.

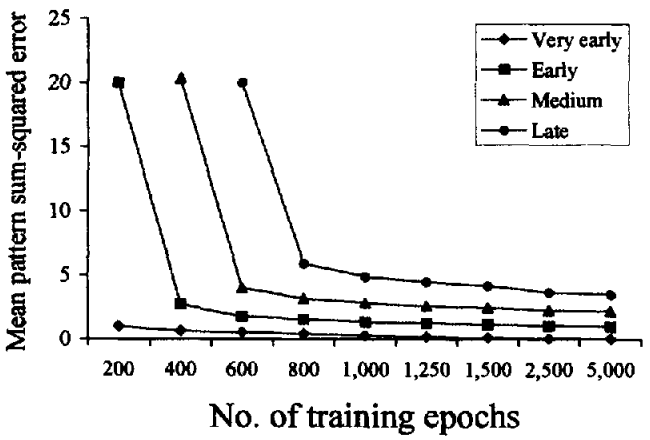

Figure 3. Effects of staggering the order of entry of pattern sets under cumulative training conditions (Simulation 5).

The analyses presented in this section included two measures of network performance in an effort to illuminate the effects of early versus late entry into training. The first was a measure of the extent to which the network learns to differentiate patterns belonging to early and late sets. This was based on analyses of the hidden unit activations to early and late patterns. The second measure was one that attempts to capture network plasticity after varying amounts of training on an early set of patterns. We show here that as training on an early set proceeds, the network becomes increasingly committed to representing those patterns and, as a result, less and less able to assimilate new, late patterns. It is this decline in plasticity with training that, we believe, underlies the effects of age (or order) of acquisition under conditions of cumulative learning. We would argue that a similar decline in plasticity in the human lexical processing networks underlies age of acquisition effects in adult lexical processing.

Age of Acquisition and Pattern Differentiation at the Hidden Units (Simulation 6)

Two sets of 50 input–output patterns (A and B) were used for this analysis. The properties of the patterns and of the network were the same as before. The response of the network to the two sets was compared under three conditions. In Simulation 6.1, the network was trained on Set A alone for 500 epochs, at which point its response to Sets A (trained) and B (untrained) was assessed. Simulation 6.2 resembled Simulation 1 in that the patterns of Set A (early) were trained from the outset, whereas the patterns of Set B (late) were entered at 250 epochs. Both sets were then trained to 500 epochs. In Simulation 6.3, both sets were trained together from the outset, and the response of the network was determined at 500 epochs.

After 500 epochs of training, the levels of activation were determined for each of the 50 intermediate (hidden) units in the network for each of the Set A and Set B patterns in each of the three simulations. The similarity of the intermediate unit activations generated by all 100 patterns was then compared for each simulation by computing the Euclidean distances between them. This yielded three similarity matrices for the patterns, one for each simulation. A multidimensional scaling procedure was then applied to each matrix to express its similarity structure in two-dimensional form. The result of this procedure is a graph of the form shown in Figure 4, in which clustering of points around the origin indicates a lack of differentiation between patterns (i.e., poor learning), whereas optimal learning results in a wide circle of points denoting well-differentiated patterns.

Figure 4A, Figure 4B, and Figure 4C show the results of the multidimensional scaling analysis of intermediate unit activations for Sets A and B at 500 epochs for Simulations 6.1, 6.2, and 6.3, respectively. In Simulation 6.1 (Figure 4A), the network was trained on Set A only for 500 epochs, at which point its responses to Sets A and B were assessed. As would be expected, the trained patterns of Set A (solid diamonds) achieved good differentiation (mean distance to origin: 1.81), whereas the untrained patterns of Set B (open squares) remained clustered around the origin (mean distance to origin: 0.79). In Simulation 6.2 (Figure 4B), the early

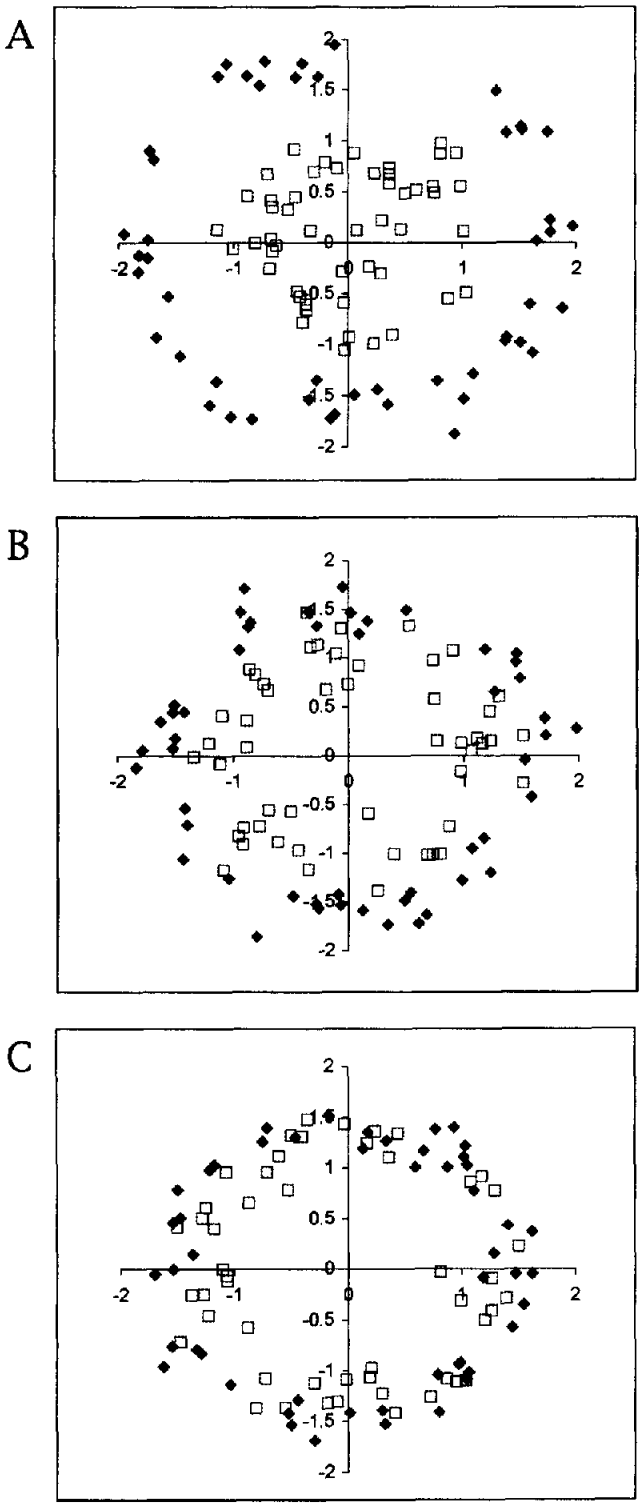

Figure 4. Differentiation of early and late patterns. Shown are the results of a multidimensional scaling analysis of intermediate unit activations for two sets of patterns, A and B, under three training conditions: trained on Set A only for 500 epochs (Simulation 6.1; A), trained on Set A for 250 epochs and then on Sets A and B to 500 epochs (Simulation 6.2; B), or trained on Sets A and B together for 500 epochs (Simulation 6.3; C). Set A = solid diamonds; Set B = open squares.

patterns of Set A again spread out to form a well-differentiated ring of points with a mean distance to origin of 1.62. In contrast, the late patterns of Set B (introduced after 250 epochs) tended to form a ring within that created by the early set with a mean distance to origin of 1.14. In other words, the late patterns struggled to achieve the same degree of differentiation as the early patterns. The difference between the distances from the origin of the early and late patterns was significant, $t(98) = 12.1$, and at 500 epochs the mean sum-squared error on the output layer was also significantly lower for the early set (0.5) than for the late set (2.5), $t(98) = 10.3$. Finally, in Simulation 6.3 (Figure 4C), in which both sets were trained together from the outset, the points for the two sets formed overlapping rings with similar mean distances from the origin (Set A = 1.49, Set B = 1.31).

Age of Acquisition and Network Plasticity

In Simulation 4, early patterns fared better than late patterns, even when the late patterns were trained at a higher frequency to the point where both sets had been trained equally often. Thus, there is more to age of acquisition effects in the network than simple cumulative frequency of training: The point at which patterns are entered into training is critical, so that if a set of additional patterns is entered after extensive training on other patterns, the network struggles to learn those new patterns. Thus, as the model learns the appropriate mappings for one set of patterns, it becomes increasingly stable and rigid, showing a resultant decrease in its capacity to assimilate new patterns.

This reduced plasticity in the network can be traced directly to the nature of learning in such networks. Distributed processing networks trained by back-propagation reduce the error between the observed pattern of activation and the target pattern by gradually changing the weights within the network. This is achieved by computing the activations of all units in the network, passing activation forward from the input units to the hidden layer and then to the output layer. During learning, the network computes an error signal for each unit at the output layer and propagates this error back through the network, first to the hidden units and then to the input layer. Over time, this iterative process gradually adjusts the weights between the layers such that the output of the model comes closer and closer to the ideal target.

When back-propagation is combined with a logistic activation function for each unit, it can be shown that the error signal for each unit in the network is proportional to the function $a(1 - a)$, where a is the activation of that unit (see McClelland & Rumelhart, 1988, pp. 126–132; Rumelhart, Hinton, & Williams, 1986). With random initial weights, the activations of all units within the network begin at a value close to 0.5. The consequence of this is that, at the start of training, the rate of weight change is at its maximum. As training proceeds, unit activations for particular patterns gradually move away from 0.5 toward either 0 or 1, and the rate of change gradually reduces. Of course, if the unit value reached 0 or 1, no error would be back-propagated through the network, and no further learning would occur. Thus, $a(1 - a)$ reaches its maximum at $a = 0.5$ and approaches its minimum as a approaches 0 or 1. Because the amount of change in a given weight is proportional to $a(1 - a)$, weights will be changed most for those units that are nearest to their midrange. Weights will be changed least for units whose activation functions have approached 0 or 1 and have thus

become effectively committed to being either on or off (see McClelland & Rumelhart, 1988, p. 132). As learning proceeds, units shift in one direction or the other, and the amount of weight change gradually reduces. When late patterns are entered into the training process, the rate of change has already fallen from its maximal value, and so those patterns are always at a disadvantage throughout any amount of additional training.

Plasticity and training. Further analyses of Simulation 6.1 illustrate this more concretely. In that simulation, Set A patterns were trained alone up to 500 epochs. Set B patterns remained untrained and therefore undifferentiated when tested at 500 epochs (see Figure 4A). We were interested in analyzing the extent to which the network would have been receptive to Set B patterns at different stages throughout its training on Set A. We computed a measure of network plasticity, or receptivity, by testing the hidden unit activations generated by the untrained patterns of Set B after no training on Set A, 50 epochs of training on Set A, 100 epochs of training on Set A, and so on up to 500 epochs of training on Set A. The activation value on each hidden unit for each of the Set B patterns was entered into the function $a(1 - a)$, and resultant figures were then averaged across the items to produce a measure of the plasticity of the hidden units. Figure 5 shows the mean plasticity values for the hidden units when responding to the untrained patterns of Set B as training proceeded on Set A. The figure shows how the plasticity of the network—that is, its responsiveness to untrained Set B when first presented—declined as the network learned to differentiate and represent the Set A patterns. A one-way analysis of variance confirmed the reduction in plasticity during training, $F(1, 450) = 205.9$, $MSE = 0.0013$.

Effects of different delays between early and late patterns on error, differentiation, and plasticity. The analysis of network plasticity suggests that the age of acquisition effect in the differential error scores on the output layer of the network should be directly related to the delay before the late patterns are entered. We compared three simulations that varied the delay between early and late patterns to assess network plasticity for late patterns as a function of delay or the degree of head start given to early items. Simulation 6.2, like Simulation 1, used a 250-epoch delay between early and late patterns, which were then trained together to 500 epochs. This was compared with a simulation in which the late patterns were entered into training after 100 epochs (Simulation 6.4) and a simulation in which late patterns were entered into training after just 50 epochs (Simulation 6.5). The mean sum-squared error for the early and late patterns at 500 epochs in each simulation is shown in Figure 6A. Analysis of variance revealed an effect of age of acquisition, $F(1, 98) = 84.4$, $MSE = 103.3$, with early patterns producing lower error scores than later ones, and an effect of delay, $F(2, 196) = 80.1$, $MSE = 18.1$, such that error across the early and late sets combined was less for shorter delays than for longer delays. But there was also a significant interaction between the variables, $F(2, 196) = 63.5$, $MSE = 14.3$, such that the difference between early and late sets increased with the size of the delay. The longer the delay in introducing the late set, the higher the error on that set after further training.

Figure 6B shows the results of the differentiation analysis on early and late patterns with the three different delays. Rather than showing the results as a two-dimensional scatterplot (as in Figure 4), we show the mean distance to the origin for the early and late patterns at each delay. As would now be predicted, the pattern differentiation (the diameter of the circle of points) was significantly greater for the early than the late sets, $F(1, 98) = 123.0$, $MSE = 10.5$. Although the main effect of delay was not significant, $F(2, 196) < 1$, there was a significant interaction between delay and age of acquisition, $F(2, 196) = 24.2$, $MSE = 0.18$, such that the difference between early and late patterns increased with the size of the delay. As delay increased, the capacity of the network to differentiate the late patterns decreased.

Figure 6C is based on Simulation 6.2 in which the entry of Set B was delayed until 250 epochs. It shows the plasticity of the network—that is, the mean of the hidden unit $a(1 - a)$ values—for Set B patterns after 50, 100, and 250 epochs of training on Set A (i.e., at the points where the Set B patterns were entered into training in Simulations 6.5, 6.4, and 6.2, respectively). Network plasticity declined with training because the network became increasingly committed to differentiating patterns trained from the outset. The result was that the network became less and less able to achieve optimal differentiation of the later patterns (Figure 6B) and error for the late patterns became increasingly great, even after training on both sets (Figure 6A).

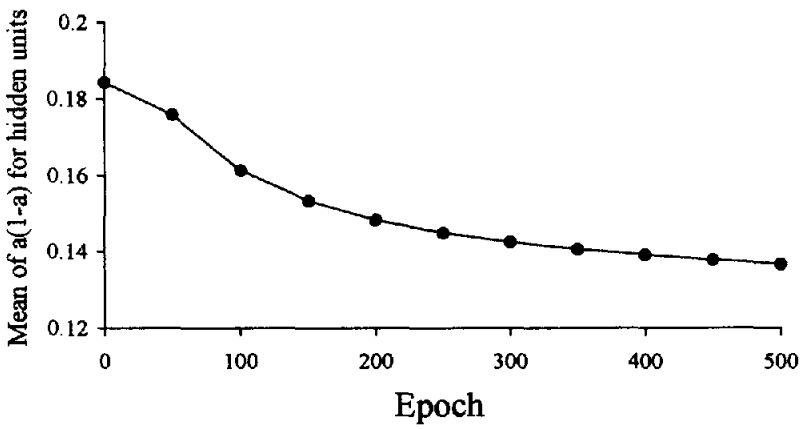

Figure 5. Mean plasticity of hidden units for the untrained patterns of Set B as a function of the extent of training on Set A (Simulation 6.1; see text for details).

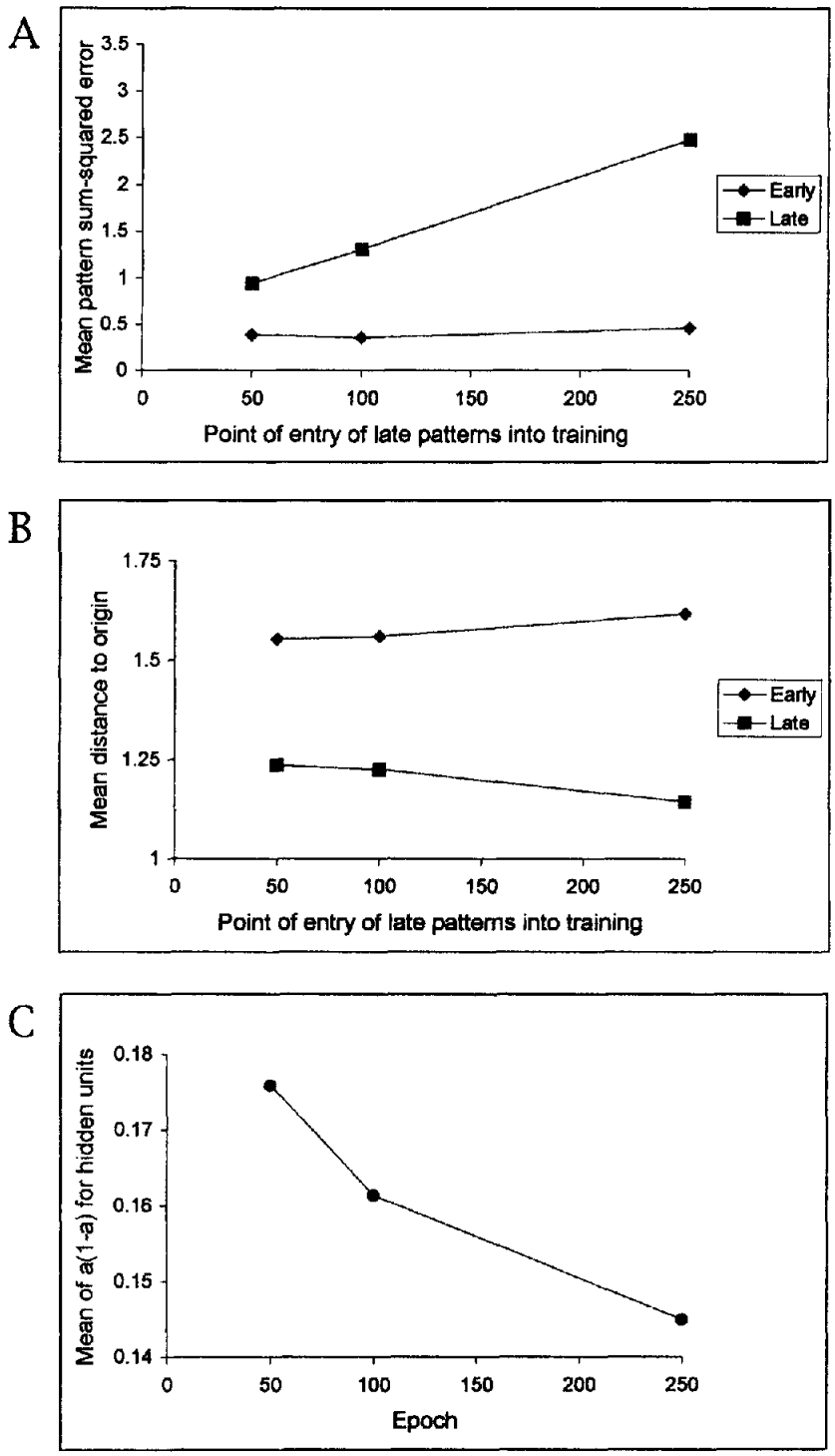

Figure 6. Effects of point of introduction of late patterns into training on error (A), pattern differentiation (B), and network plasticity (C). Values are based on Simulations 6.2, 6.4, and 6.5 (see text for details).

Simulation 7: Sparseness of Coding

Simulations 7 to 10 explored the consequences for the network age of acquisition effect of varying certain characteristics of the patterns and the learning procedure. The characteristics concerned were the sparseness of the patterns (Simulation 7), the size of the early and late pattern sets (vocabularies) that the network was required to learn (Simulation 8), the learning rate used in the model

(Simulation 9), and the form of the learning algorithm (Simulation 10). Consideration of the effects of introducing weight decay into the system (Simulations 14 and 15) is deferred until after the presentation of simulations concerned with the joint effects on learning of age of acquisition and frequency of training (Simulations 11 to 13).

The patterns used in the simulations reported thus far all involved relatively sparse coding, with each pattern being distributed across 100 input or output units but with only 20% of units on average being on (i.e., set from 0 to 1). Simulation 7 examined the consequences of varying the sparseness of coding on the age of acquisition effect. In each simulation, 100 early patterns were trained from the outset, and 100 late patterns were entered into training after 250 epochs. All patterns were presented once per epoch. Performance of the network was examined after 500 epochs. Sparseness of coding was varied systematically. The sparsest pattern sets, used in Simulation 7.1, were those in which input and output units were set from 0 to 1 with a probability of .1, meaning that the average 100-unit pattern had only 10 units set to 1. The probability of switching was then increased by increments of .1, from .2 (Simulation 7.2) up to .9 (Simulation 7.9).

Figure 7 shows the performance of the network on the early and late sets at 500 epochs for each of the nine levels of sparseness. An analysis of variance was carried out on the error scores for each pattern after 500 epochs of training, with age of acquisition (two levels) and sparseness of coding (nine levels) as variables. This analysis revealed significant effects of age of acquisition, $F(1, 1782) = 2{,}804.2$, $MSE = 4{,}968.0$, and sparseness, $F(1, 1782) = 3.7$, $MSE = 6.5$, but no interaction, $F(1, 1782) < 1$. Overall, learning was worst for patterns in which the probability of a unit being switched from 0 to 1 was .5. Learning was best when the patterns were effectively very sparse, that is, at .1 or .9 (note that for Simulation 7.9, the patterns were differentiated by the small number of zeros rather than the small number of ones, as for Simulation 7.1). There was a clear age of acquisition effect at all levels of sparseness of coding.

Simulation 8: Effects of Varying Vocabulary Size

The consequences of changing vocabulary size were investigated in Simulation 8. The same network and pattern characteristics were used as in Simulation 1, with early patterns being trained from the outset and late patterns being introduced after 250 epochs. Training was continued to 500 epochs. The smallest vocabulary used had 50 patterns divided into 25 early and 25 late patterns (Simulation 8.1). The same two sets of 25 patterns were included as subsets in larger vocabularies that had 100, 150, or 200 patterns, half early and half late in each case (Simulations 8.2 to 8.4, the last being a replication of Simulation 1).

Figure 8 shows the performance of the network on the early and late sets at 500 epochs for each vocabulary size. An analysis of variance was carried out on the error scores after 500 epochs of the subset of 25 early and 25 late patterns shared across the simulations; age of acquisition (two levels) and vocabulary size (four levels) were variables. There were significant effects of age of acquisition, $F(1, 48) = 170.9$, $MSE = 305.62$, with error being less to early than late patterns, and vocabulary size, $F(3, 144) = 34.6$, $MSE = 24.56$, with error at 500 epochs increasing with the total number of patterns to be learned. The interaction between age of acquisition and vocabulary size was also significant, $F(3, 144) = 14.9$, $MSE = 10.62$, reflecting the fact that the difference between early and late patterns increased with increasing vocabulary size. This was particularly reflected in worse performance on the late patterns as vocabulary size increased.

This effect can also be shown to be related to reducing plasticity with increased training. Hidden unit values for the late patterns were recorded at 250 epochs (the point at which these patterns were entered into training) for the four different vocabulary sizes used in the simulation. The average hidden unit plasticity was calculated in the same way as before. The values are plotted in Figure 9, where it can be seen that the more early patterns the network is exposed to, the less plasticity it shows for patterns introduced later. The decline in plasticity as a function of early vocabulary size was significant, $F(3, 147) = 24.7$, $MSE = 0.0087$.

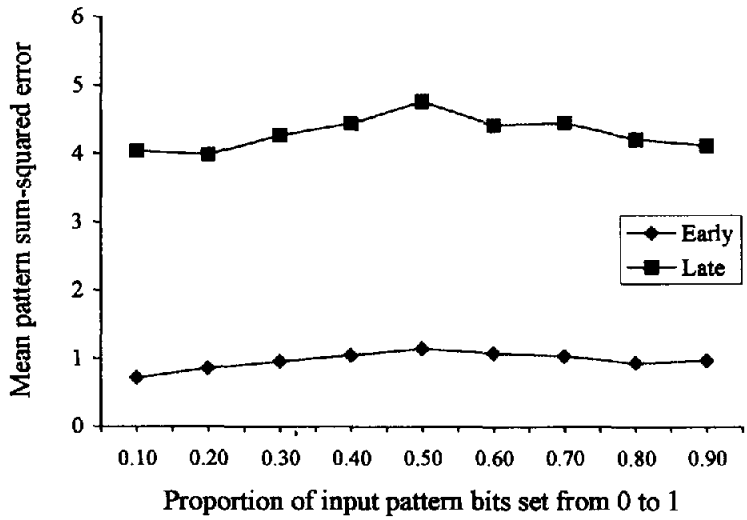

Figure 7. Effects of different degrees of sparseness of coding on early and late patterns (Simulation 7).

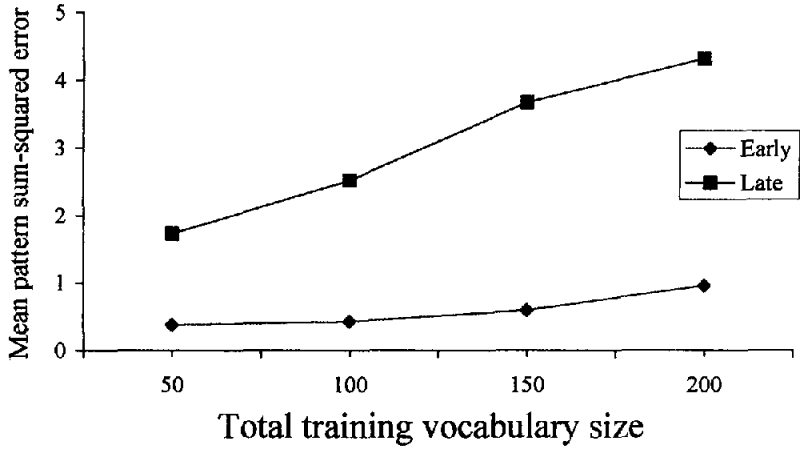

Figure 8. Effects of vocabulary size on early and late patterns (Simulation 8).

Simulation 9: Effects of Varying Learning Rate

In all of the simulations reported thus far, the learning rate parameter of the model was fixed at a value of 0.05. The learning rate parameter governs how much the weight values in the network are adjusted after each pattern is presented during training. A smaller value means that the weights are changed by a lesser amount each time they are updated.

Simulation 9 used the same 100 early and 100 late patterns as in Simulation 1, with the late set being entered after 250 epochs and network performance being tested at 500 epochs. Values of the learning rate parameter were varied from 0.01 to 1.0. Analysis of variance on the error data at 500 epochs showed significant effects of age of acquisition, $F(1, 198) = 499.9$, $MSE = 10{,}030.5$, and learning rate, $F(14, 2772) = 191.9$, $MSE = 173.8$, together with a significant interaction between the two, $F(14, 2772) = 9.9$, $MSE = 8.95$. Figure 10 shows that increasing learning rate from the very low level of 0.01 to 0.1 improved the efficiency with which the network could learn both sets of patterns by 500 epochs. Thus, error on the early patterns declined from 4.25 to 0.42 over that range of values of learning rate, whereas error on the late patterns declined from 7.28 to 3.87. As learning rate increased further from 0.1 to 0.5, error on the early patterns continued to decline down to 0.28, but error on the late patterns increased to 4.25. Over this range, better learning of the early set over the first 250 epochs came at the cost of a reduced capacity to learn the late set when introduced into training. Finally, learning rates above 0.5 were increasingly deleterious to the learning of early and late sets alike. (It is well known that if learning rate is set too high, learning by back-propagation becomes inefficient.) Importantly, variations in learning rate had only minor consequences for the age of acquisition effect over the range of learning rates that brought about effective learning of both pattern sets. Our standard value for learning rate (0.05) falls within this range.

Simulation 10: Learning With Quickprop

With standard back-propagation, the activation functions of units can go all the way from their initial value of about 0.5 to their maximum or minimum value of 1 or 0. Once an activation function

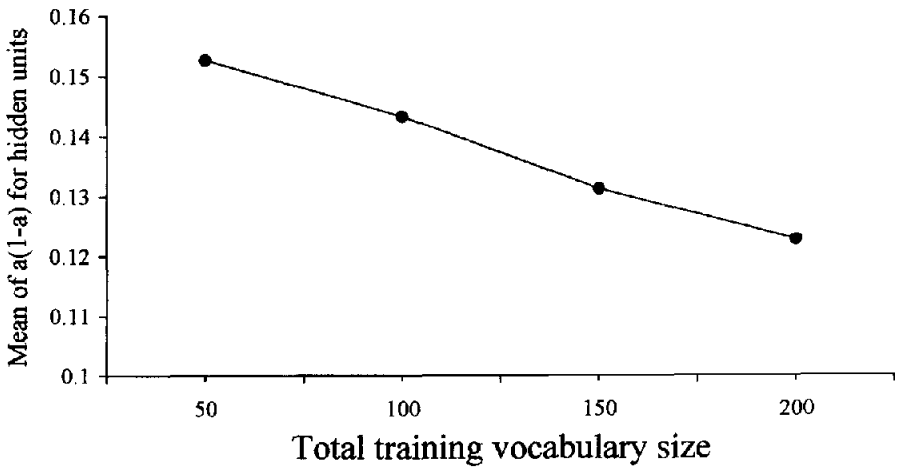

Figure 9. Effects of total vocabulary size on network plasticity (based on Simulation 8).

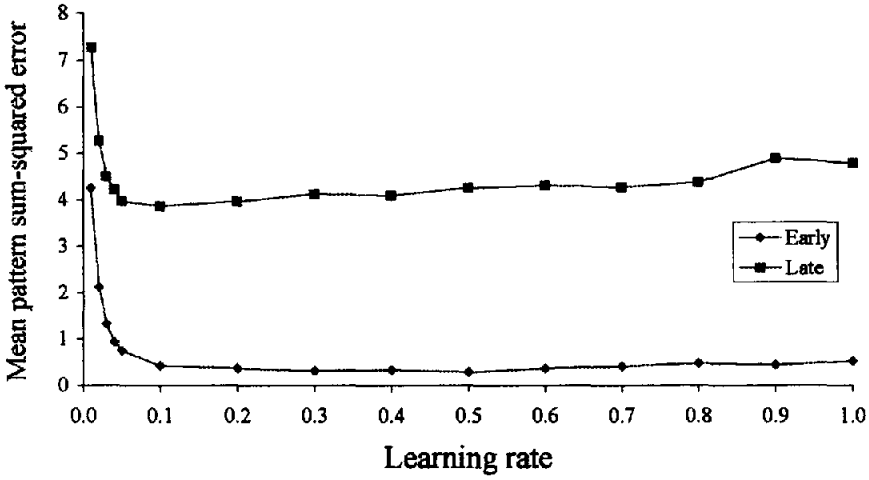

Figure 10. Effects of varying learning rate on network performance for early and late items (Simulation 9).

has approached 1 or 0, it becomes very difficult for further training to pull it back again. Hence, back-propagation networks show a tendency toward progressive entrenchment and rigidification as learning proceeds.

Fahlman (1989) developed a variant of back-propagation called Quickprop, which reduces this tendency by preventing plasticity from ever falling to extremely low levels. This was done by adding a small constant to the expression $a(1 - a)$ when calculating the derivatives of the output and hidden units. This modification of the back-propagation algorithm was shown to improve learning of sets of associations in which some items were intrinsically easier to learn than others, although it has also been reported to increase the tendency of a network to become stuck in local minima in the course of learning (Veitch & Holmes, 1991).

Simulation 10 repeated Simulation 1, but with the Quickprop algorithm in place of standard back-propagation. The results are shown in Figure 11, where the x-axis shows the parameter *qpoffset*, which determines the value of the constant added to $a(1 - a)$. Analysis of variance showed significant effects of age of acquisition, $F(1, 198) = 582.0$, $MSE = 8{,}204.5$, and qpoffset, $F(9, 1782) = 356.0$, $MSE = 347.3$, together with a significant interaction between the two variables, $F(9, 1782) = 46.7$, $MSE = 45.6$. Figure 11 shows, however, that the age of acquisition effect was relatively constant across those values of qpoffset over which effective learning occurred within 500 epochs. (Note that learning with qpoffset set at 0 is learning with standard back-propagation, so these data points are replications of Simulation 1.)

One might suspect that effects of variations in the learning algorithm could reveal themselves over longer periods of training. To check this, we allowed Simulation 10 to continue running to 10,000 epochs with qpoffset values of 0, 0.02, and 0.04. Learning was slightly faster with Quickprop than with standard back-propagation, but performance on the early and late sets at 10,000 epochs was virtually the same, as was the difference between performance on early and late pattern sets. Hence, Quickprop did not turn out to influence the age of acquisition effect to any

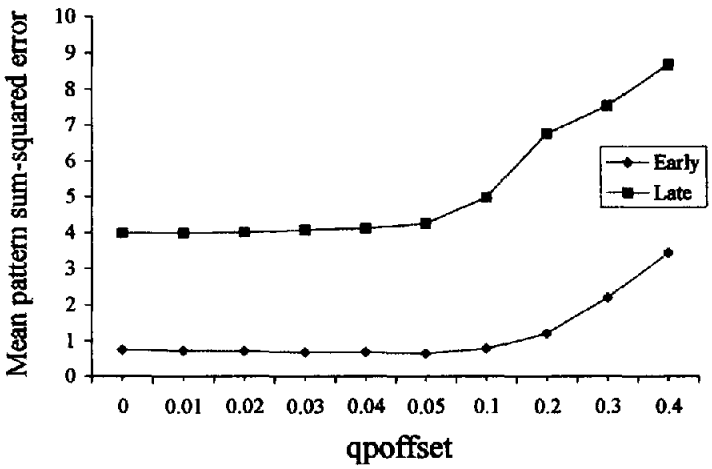

Figure 11. Performance on early and late patterns using the Quickprop learning algorithm (Simulation 10; see text for details).

significant extent. This implies that the effect does not rely on hidden unit activations reaching extreme values close to 0 or 1. Note, however, that hidden unit activation follows a sigmoid function, so early movements away from a midvalue of approximately 0.5 are greater than later movements that occur when the function has moved toward 0 or 1. Progressively later training has progressively smaller effects, and hence there is a loss of plasticity long before minimum or maximum values are reached. Note also that training of early and late patterns is constantly interleaved so that once the late patterns have entered into training, the early patterns are presented along with them in every epoch. Quickprop prevents plasticity from dropping to 0, leaving the model to back-propagate error forever; it will continue to back-propagate error for both sets of patterns, however, so the difference inherited by the early patterns remains intact.

Simulation 11: Combining Age of Acquisition and Frequency

The simulations reported thus far have shown that age of acquisition effects are perfectly compatible with connectionist networks trained by back-propagation. Effects of different points of entry into training cannot be reduced to simple differences in cumulative frequency (Simulations 3 and 4). But frequency of presentation does, in fact, have an effect on network performance: The more often a network is trained on a particular pattern (or association between patterns), the stronger the resulting representation becomes (e.g., Plaut et al., 1996; Seidenberg & McClelland, 1989).

There has been some debate in the experimental literature over whether frequency effects in word recognition and production actually survive the experimental control of age of acquisition or its factoring out by statistical means in regression analyses (e.g., Gerhand & Barry, 1998; Gilhooly & Watson, 1981; Morrison & Ellis, 1995). There seems to be a growing consensus, though, that both the age at which a word is learned and the frequency with which it is subsequently encountered affect reaction time in adult object naming (e.g., Barry et al., 1997; Ellis & Morrison, 1998), word naming (Gerhand & Barry, 1998, 1999a; Morrison & Ellis, 2000), and visual lexical decision (Butler & Hains, 1979; Morrison & Ellis, 1995; Nagy, Anderson, Schommer, Scott, & Stallman, 1989). The literature on human lexical processing is, however, ambiguous as to the nature of the interaction between these two factors. Some studies have reported stronger frequency effects for early than late acquired words (Barry et al., 1997; Gerhand & Barry, 1999a), whereas others have reported additive effects (Gerhand & Barry, 1998).

Simulations 11.1 and 11.2 used the same network and the same 200 input–output patterns as in Simulation 1. Training was always cumulative and interleaved. To manipulate frequency of presentation, we randomly divided the sets of 100 early and 100 late patterns into two subsets of 25 and 75 patterns each. In recognition of the fact that there are more low-frequency than high-frequency words in the language, the subsets of 25 patterns were trained with higher frequencies than the subsets of 75 patterns. The early sets were trained from the outset, whereas the late sets were entered into training after 750 epochs. Two differences in frequency of training were compared. In Simulation 11.1 the high-frequency patterns were presented 3 times per epoch, whereas in Simulation 11.2 the high-frequency patterns were presented 10 times per epoch. The low-frequency patterns were presented once per epoch in each case. Training was once again by standard back-propagation. The network stabilized by approximately 3,000 epochs, and performance on the four pattern sets in each simulation was assessed after 5,000 epochs of training.

Figure 12A and Figure 12B show the mean pattern sum-squared error scores at 5,000 epochs for the early and late high- and low-frequency sets of Simulations 11.1 and 11.2, respectively. These error scores from the two simulations were entered into an analysis of variance with age of acquisition (early or late entry), pattern frequency (low or high frequency), and frequency difference (3 or 10 presentations per epoch for the high-frequency patterns) as variables. The results showed a significant effect of age of acquisition, $F(1, 196) = 225.2$, $MSE = 684.1$, with error being less to early than late patterns, and a significant effect of pattern frequency, $F(1, 196) = 15.3$, $MSE = 46.3$, with error being less to high- than low-frequency patterns. The main effect of frequency difference was not significant, $F(1, 196) < 1$, meaning that overall learning by 5,000 epochs was similar whether the frequency differential was 3:1 or 10:1. The interaction between age of acquisition and pattern frequency was not significant, $F(1, 196) = 1.4$, $MSE = 4.2$, $p = .24$, but the three-way interaction among age of acquisition, pattern frequency, and frequency difference approached significance, $F(1, 196) = 3.4$, $MSE = 4.0$, $p = .07$. As Figure 12A suggests, there was no interaction between frequency and age of acquisition when the high-frequency patterns were presented only 3 times per epoch, $F(1, 196) < 1$, but the same interaction depicted in Figure 12B approached significance when the high-frequency patterns were presented 10 times per epoch, $F(1, 196) = 3.3$, $MSE = 8.2$, $p = .07$. This result implies that the interactivity of age of acquisition and frequency in human lexical processing experiments may depend on the strength of the frequency manipulation (and also, perhaps, on the strength of the age of acquisition manipulation).

By 5,000 epochs, the late, high-frequency patterns in Simulation 11.2 (Figure 12B) had each been presented 42,500 times in training. The early, low-frequency patterns had been presented only 5,000 times each by the same point. Nevertheless, the head start of 750 epochs given to the early, low-frequency patterns meant that they continued to outperform the late, high-frequency patterns at 5,000 epochs, even though the late, high-frequency patterns had been presented 8.5 times as often.

Simulation 12: From High to Low Frequency—Simulating the Loss of "Childish" Words

The vast majority of words learned in childhood continue to be used in adulthood. Some words, however, occur more often in the experience of the average child than in the experience of the average adult. Examples are words such as *potty* and *nappy/diaper,* or words that are largely confined to certain nursery rhymes or children's stories (the *fleece* of Mary's little lamb or the *pail* that Jack and Jill carried up the hill). Except when adults are themselves engaged in caring for children, the frequency of such words will be lower in adulthood than it was in early childhood.

Simulation 12 examined the effect of reducing the frequency of

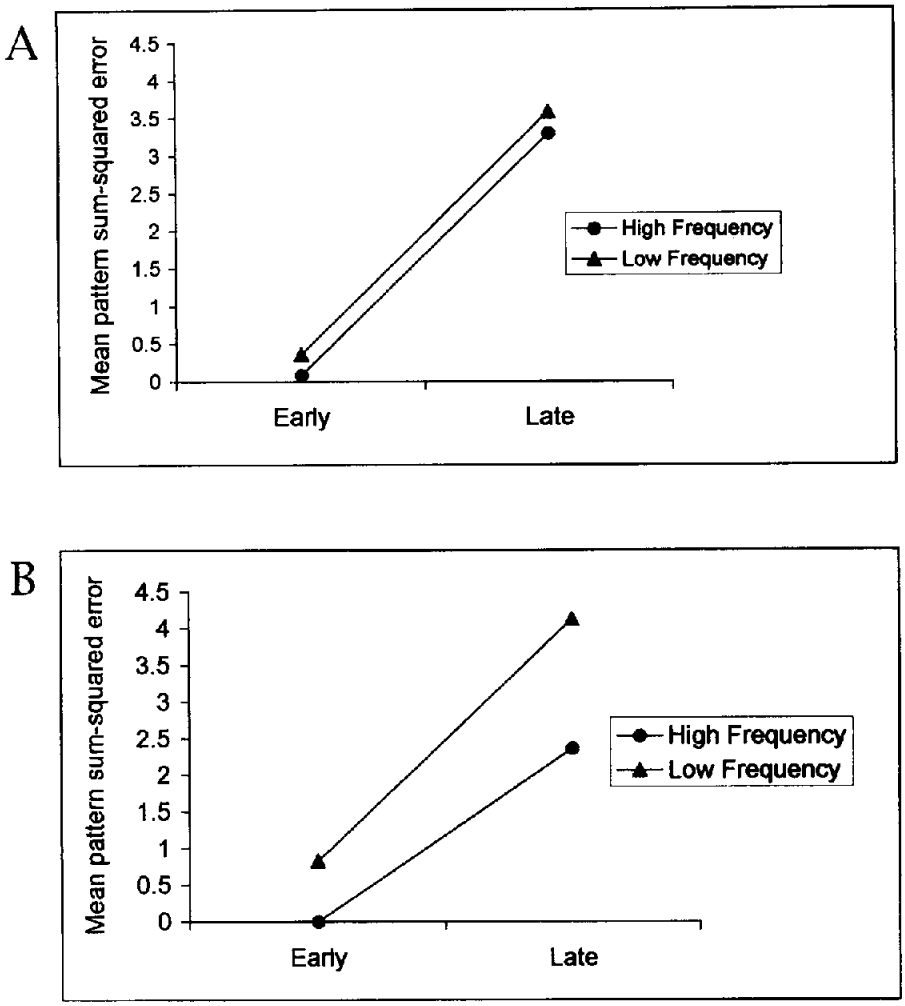

Figure 12. Performance on early and late patterns given high-frequency or low-frequency training with frequency ratios of either 3:1 (A; Simulation 11.1) or 10:1 (B; Simulation 11.2).

presentation of a set of early, high-frequency patterns after the network had learned them. As in Simulation 11.2, the network was trained from the start on two sets of early patterns, 25 with high frequency (10 presentations per epoch) and 75 with low frequency (1 presentation per epoch). At 750 epochs, two more sets of late patterns were introduced, with one set of 25 patterns being trained with a high frequency and one set of 75 patterns with a low frequency. In Simulation 12.1, presentation of the early, high-frequency set continued at its original rate of 10 per epoch, effectively a replication of Simulation 11.2. In the other versions of Simulation 12, the frequency of presentation of the early, high-frequency set was reduced at the point where the two late sets entered into training. Presentations of the early, high-frequency set were reduced from 10 to 3 per epoch in Simulation 12.2 and from 10 to 1 per epoch in Simulation 12.3. In Simulation 12.4, the frequency of presentation of the early, high-frequency set was reduced to 0 (i.e., the early, high-frequency patterns were not trained at all after 750 epochs). Training was continued to 5,000 epochs in all simulations.

Figure 13 shows the mean pattern sum-squared error at 5,000 epochs for each pattern set in each simulation. Simulation 12.1 replicated the results of Simulation 11.2, with performance on the four sets stabilizing by about 3,000 epochs, after which a constant differential between sets was maintained. Reducing presentations of the early, high-frequency set from 10 per epoch to 3 (Simulation 12.2) or even to 1 (Simulation 12.3) at the point when the late patterns were introduced (750 epochs) did not impair performance on that early set, which showed no increase in error from 750 to 5,000 epochs and continued to outperform the late, high-frequency set. Only when the early, high-frequency set stopped being trained altogether after 750 epochs did it show an increase in error (catastrophic interference).

These simulations show that once an early set of patterns has been well learned by the network, a presentation frequency much less than the original one will serve to maintain the quality of the representations. Only if the early patterns cease to be trained at all do their representations suffer. The low frequency with which most adults encounter and use words such as *potty* and *fleece* may

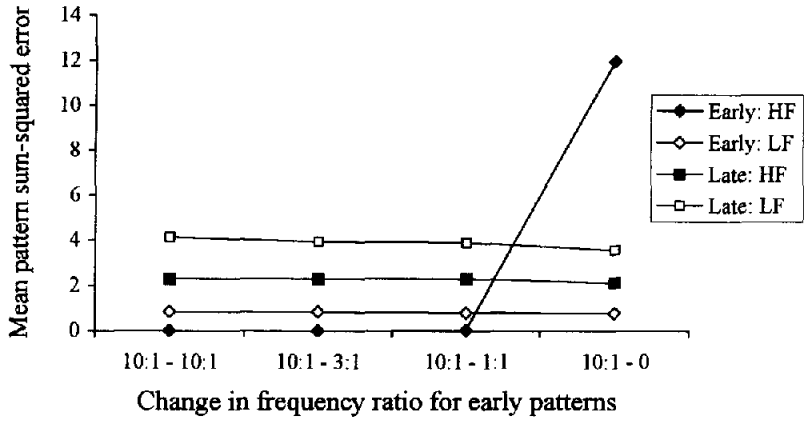

Figure 13. Effects of reducing the frequency of presentation of an early, high-frequency set after extensive training (Simulation 12). HF = high frequency; LF = low frequency.

nevertheless suffice to keep those words recognizable and accessible despite their low frequency in adulthood.

Simulation 13: Very Late Acquisition of Vocabulary

Human vocabulary acquisition slows down in adulthood but never stops completely. In recent decades, information technology, for example, has been a fruitful source of new words that may in some cases attain quite high frequencies. The word *computer* is late acquired for anyone more than 40 years of age or so, whereas the word *email* is probably late acquired for most people more than 20 years of age. Both are, however, of high frequency for many people.

Simulation 13 examined the effect of introducing a new set of very late patterns into training long after the network had stabilized on both the normal early and late sets. Simulation 11.2 was used as the basis for this (early vs. late patterns separated by a delay of 750 epochs with 10 presentations per epoch for high-frequency patterns and 1 per epoch for low-frequency patterns). A very late set of 25 new patterns was introduced at 5,000 epochs with frequencies of 1 (Simulation 13.1), 10 (Simulation 13.2), 100 (Simulation 13.3), or 1,000 (Simulation 13.4) presentations per epoch. The combined total of 225 patterns was trained for a further 5,000 epochs. One might question the psychological plausibility of the last two simulations—the frequency of very late acquired words will rarely, if ever, exceed that of earlier learned high-frequency words—but they were included here for reasons that are clear from Figure 14, which shows the error scores for the five sets of patterns at 10,000 epochs. With presentation frequencies of 1 or 10 per epoch, the very late patterns never became as well represented as earlier learned patterns. With 100 presentations per epoch, the very late patterns compared in accuracy to the late, low-frequency patterns (trained at 1 presentation per epoch). Only with 1,000 presentations per epoch did the mean error score for the very late patterns fall below that of the normal late, low-frequency patterns entered at 750 epochs and trained once per epoch thereafter. Error scores for the very late patterns were still higher than those for the

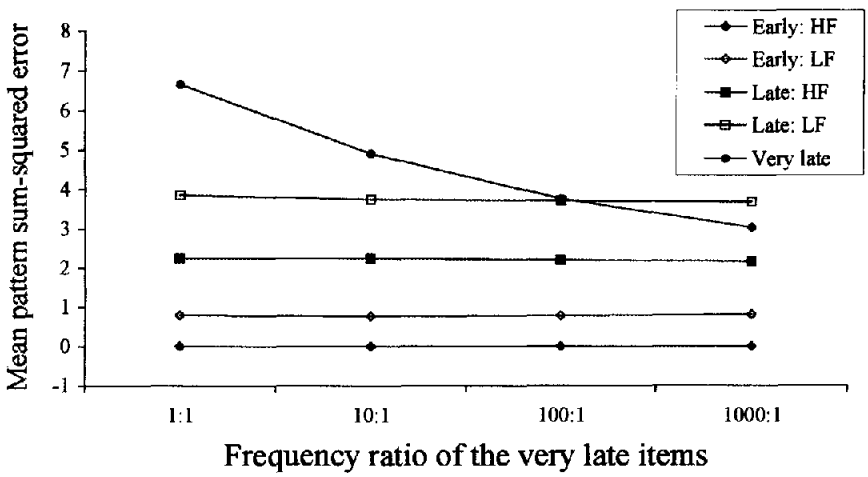

Figure 14. Effects of entering a very late set of patterns with frequencies from 1 to 1,000 presentations per epoch (Simulation 13). HF = high frequency; LF = low frequency.

late, high-frequency set or either of the early sets. Thus, once the network has become committed to one set of patterns and has been trained to the point at which it has lost much of its plasticity, considerable effort (training) is required to imprint new patterns and bring them to the condition where their representations are comparable to those established by the earlier patterns.

Simulation 14: Change and Decay—The Effects of Introducing Weight Decay

If weight decay is introduced into a connectionist network, then connection strengths gradually and spontaneously decay back to resting level unless their values are maintained by training. This means that weight decay acts against any tendency a network might have toward overlearning, that is, toward becoming excessively rigid in the course of training. One consequence of this is that the capacity of the network to generalize to new inputs is improved (Hinton, 1989). Thus, Plaut et al. (1996) introduced weight decay into some of their simulations of reading aloud to improve the ability of a network trained to associate written with spoken word forms to generalize that knowledge to pronouncing nonwords.

We have argued that age of acquisition effects arise precisely because of the tendency of networks to rigidify with training. The introduction of weight decay into our simulations might thus be expected to reduce the extent to which plasticity is lost in the course of training and hence reduce the effects of age of acquisition on learning. Simulation 14 used the same network and patterns as Simulation 1 (100 early and 100 late patterns, with the late set being introduced into training after 250 epochs). But whereas Simulation 1 (like all of the other simulations reported thus far) did not involve weight decay, Simulation 14 compared network performance on the early and late pattern sets at 500 epochs with no weight decay (Simulation 14.1, a replication of Simulation 1) and with five levels of increasing weight decay ranging from a low of 0.000001 (Simulation 14.2) to a highest level of 0.001 (Simulation 14.6). These values may appear very small, but note that weight decay is applied after each pattern has been presented. Small values of weight decay therefore have cumulative effects across an epoch. If weight decay was being applied only at the end of each epoch, these values of weight decay would be correspondingly larger.

The results are shown in Figure 15 in terms of mean pattern sum-squared error for the early and late patterns at 500 epochs with the six levels of weight decay. The main effect of weight decay was significant, $F(5, 990) = 898.0$, $MSE = 231.6$, as was the main effect of age of acquisition, $F(1, 198) = 275.7$, $MSE = 1{,}955$. The interaction between weight decay and age of acquisition was also significant, $F(5, 990) = 175.6$, $MSE = 45.3$, reflecting the reduction in the age of acquisition effect as weight decay increased. With zero or very low levels of weight decay, age of acquisition effects were large, but as weight decay increased, the age of acquisition effect began to diminish. Note, though, that with higher levels of weight decay, the network learned both early and late pattern sets less well by 500 epochs. At the higher levels, weight decay was acting against learning, and the network began to unlearn as fast as it learned. Importantly, all levels of weight decay that permitted the network to learn both the early and the late patterns by 500 epochs resulted in the network continuing to show an age of acquisition effect.

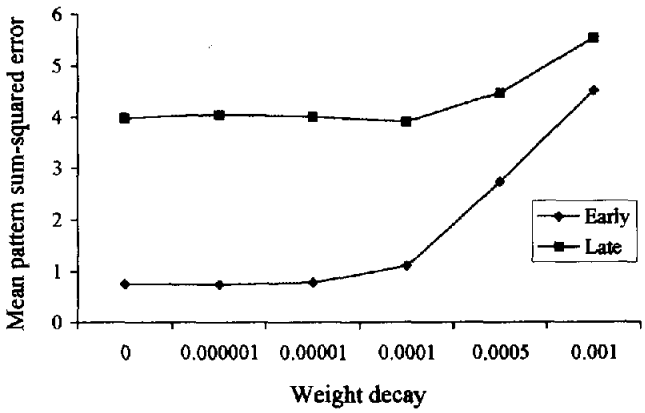

Figure 15. Effects of different levels of weight decay on network performance (based on Simulation 1).

Simulation 15: Longer Term Consequences of Weight Decay for Age of Acquisition and Frequency Effects

Simulation 11 looked at the effects of combining differences in the point at which pattern sets entered into training (age of acquisition) with differences in the frequency with which they were trained thereafter. We showed that both factors exerted an influence on the performance of the mature network. Weight decay serves to undo the effects of training unless that training is reinforced by further encounters with the patterns. It might thus be expected that low-frequency patterns would be more affected by weight decay than high-frequency patterns, because the high-frequency patterns have more opportunities to counteract the tendency of weight decay to shift weights back toward resting level. This in turn could alter the balance of power between patterns that have the benefit of early training but are trained only at low frequencies and patterns that are introduced later into training but are trained with high frequency thereafter.

Simulation 15 was a repeat of Simulation 11.2 with the addition of weight decay. One hundred early patterns were trained from the outset, and 100 late patterns were introduced after 750 epochs. The early and late sets were each divided into subsets of 75 patterns trained with low frequency (once per epoch) and 25 patterns trained with high frequency (10 times per epoch). The network was assessed after 5,000 epochs of training. There were five levels of weight decay ranging from 0 (Simulation 15.1, a replication of Simulation 11.2) to 0.0001 (Simulation 15.5). Figure 16 shows the mean error scores at 5,000 epochs for each pattern set for the different levels of weight decay. Note that the *y*-axis in Figure 16 starts below 0 to show the very low error rates to early high-frequency patterns with low levels of weight decay.

A three-way analysis of variance was carried out on the error scores at 5,000 epochs with weight decay, age of acquisition, and frequency as variables. There were significant main effects of weight decay, $F(4, 784) = 206.7$, $MSE = 242.4$; age of acquisition, $F(1, 196) = 80.7$, $MSE = 364.0$; and frequency, $F(1, 196) = 307.7$, $MSE = 1{,}387.7$. The interaction between age of acquisition

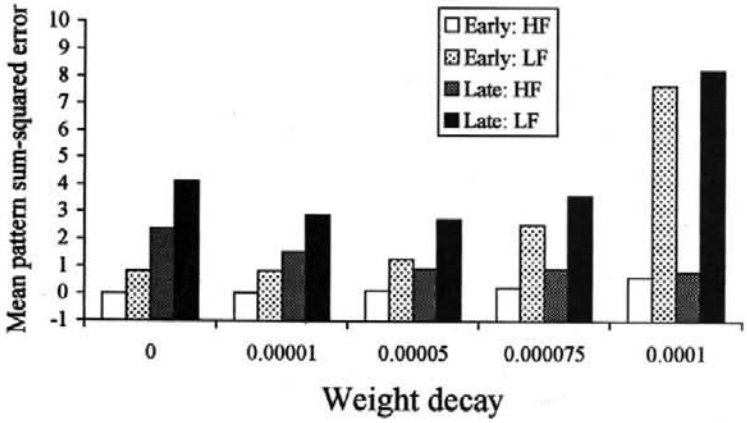

Figure 16. Effects of different levels of weight decay on pattern sets varying in age of acquisition and frequency. The network was trained to 5,000 epochs (based on Simulation 11.2). HF = high frequency; LF = low frequency.

and frequency approached significance, $F(1, 196) = 3.5$, $MSE = 15.0$, $p = .06$, indicating that frequency effects tended overall to be greater for late than early acquired patterns. There were also significant first-order interactions between weight decay and age of acquisition, $F(4, 784) = 28.8$, $MSE = 33.7$, and between weight decay and frequency, $F(4, 784) = 215.1$, $MSE = 252.2$. The second-order interaction among weight decay, frequency, and age of acquisition was not significant, $F(4, 784) < 1$. The interactions of weight decay with age of acquisition and frequency show that as weight decay increased, the relative influences of age of acquisition and frequency changed. With low levels of decay, point of entry into training was very important (as in Simulation 11); with higher levels of decay, frequency of training became more important (because there was a need for continued training to battle against the tendency of weights to revert to their initial levels unless constantly refreshed). With the highest level of weight decay (0.0001), the network failed to learn either of the sets of low-frequency patterns by 5,000 epochs.

Separate analyses of variance were carried out on the error scores at 5,000 epochs for the different levels of weight decay that permitted the network to learn all four sets to a reasonable extent by 5,000 epochs (0 to 0.000075). Even though the balance of influence between them changed with the addition of weight decay, there were significant effects of both age of acquisition and frequency on error scores at all four levels of decay.

Simulation 16: Lesioning the Network

Age of acquisition has been shown to affect the accuracy with which normal elderly people, patients with poststroke aphasia, and patients with semantic dementia can name objects. These groups all fail to name objects with late acquired names that young, normal adults name correctly (Ellis et al., 1996; Hirsh & Ellis, 1994; Hirsh & Funnell, 1995; Hodgson & Ellis, 1998; Lambon Ralph et al., 1998; Nickels & Howard, 1995).

We studied the effect of "lesioning" a trained network by resetting a proportion of connections to and from the hidden units to zero. The starting point was the network from Simulation 3 that had been trained on 100 early and 100 late patterns up to 5,000 epochs, with the late patterns being introduced into training after 250 epochs. The network was lesioned by reducing the connection strengths (weights) to zero with probabilities of .05, .10, .15, and .20. The results are shown in Figure 17. Error increased as the network was lesioned more. If an arbitrary threshold is set, above which the network was deemed not to be able to generate the required output pattern for a given input pattern, then it is clear from Figure 17 that a damaged network would always be better able to generate correct responses to early than late items, as humans can.

General Discussion

Simulations 1 and 2 showed that analogies drawn in the past between age of acquisition effects in adults and catastrophic interference in connectionist networks were misguided. Interference with consequent loss of representations occurs when one set of items entirely replaces another in training (Simulation 2) or when items cease to be presented in training (Simulation 12.4). But when items introduced early into training are then joined by later sets that are trained alongside them in a cumulative and interleaved manner, the performance of the network continues to favor the early set (Simulations 1 and 3). The advantage for early acquired patterns cannot be explained simply in terms of differences between early and late sets in cumulative frequency of training (Simulations 4 and 11).

The age of acquisition effect does not simply reflect an advantage for items trained from the outset over other items: If the point of entry into training of different pattern sets is staggered, then performance of the network after extensive training reflects the order of entry of the patterns in training (Simulation 5). If differences in the order of entry of pattern sets into training are combined with differences in the frequency with which patterns are trained, then the final performance of the network reflects both factors (Simulation 6).

Analysis of network performance provides some insights into why it is that the network struggles to learn patterns introduced into training once it has been trained on earlier patterns. If the network is to learn the association between input and output

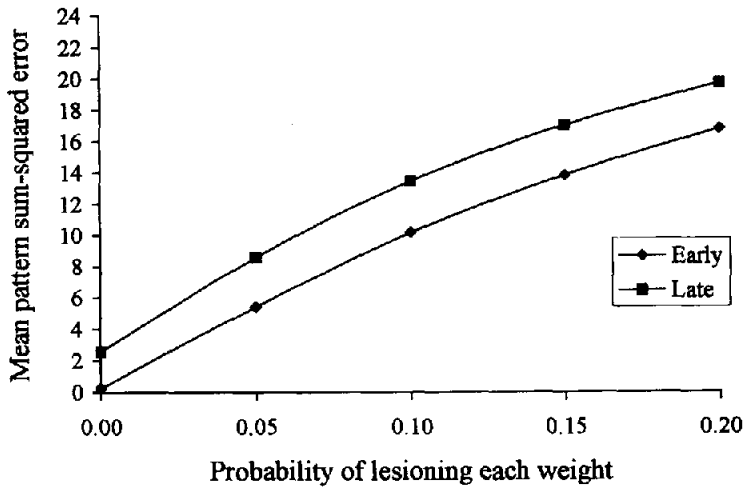

Figure 17. Effects on error scores for early and late patterns of lesioning a trained network by resetting weights to zero with probabilities ranging from .05 to .20 (Simulation 16).

patterns, it has to adjust the weights between units in the different layers. This tends to shift the activations of units in the intermediate layer away from the starting value of 0.5 toward either 0 or 1. Early training moves the function away from the initial value more than later training can. The result is that the intermediate units become progressively committed to achieving maximum differentiation between the patterns introduced at the outset of training. This commitment comes at a cost, which is that the structure of the network loses plasticity and becomes less and less efficient at learning and representing associations introduced later (Simulation 6). Loss of plasticity does not, however, require that unit activations reach maximum or minimum values: Use of a modified version of back-propagation that prevents this (Quickprop) did not significantly reduce the difference in the network's response to early and late pattern sets (Simulation 10). The fact that early movements of the activation function away from resting level are greater than later movements appears sufficient to bestow an advantage on early patterns.

Further simulations with standard back-propagation showed that age of acquisition effects are characteristic of patterns with varying degrees of sparseness of coding (Simulation 7), increase with increasing vocabulary size (although they are present for all vocabulary sizes; Simulation 8), and occur for all values of the learning rate parameter sufficient to induce learning in the network (Simulation 9). The point at which items are entered into training (age of acquisition) is not, of course, the only factor that affects network performance. The frequency with which patterns are trained also influences network structure. Simulation 11 showed that the contributions of age of acquisition and frequency can appear additive or interactive depending on the strength of the frequency manipulation (and, presumably, the strength of the age of acquisition manipulation). Once the model had been trained, the frequency with which early patterns that were originally of high frequency continued to be presented to the mature network could be reduced considerably without any increase in the error associated with those patterns. In the absence of weight decay, it was only when training ceased altogether that performance on early, high-frequency patterns declined (Simulation 12).

It is hard to teach old networks new tricks. Patterns entered into training after the network had more or less stabilized needed to be presented at very high frequencies indeed before their representations became comparable to those of earlier acquired patterns (Simulation 13). In human terms, this predicts that words acquired in adulthood will struggle to develop representations comparable to those of words learned in childhood. Examples might be words entering the language as a result of technological developments or words associated with particular adult occupations.

The behavior of the network was altered to a degree if it was trained with an element of weight decay. When weight decay is present, connections that are not continually refreshed decay back toward their starting levels. Simulation 14 showed a tendency for weight decay to reduce (but not eliminate) age of acquisition effects. Simulation 15 examined the way in which the introduction of weight decay alters the balance of influence between age of acquisition and frequency. The presence of weight decay means that training has to combat the tendency for weights to return to resting levels through decay. High-frequency training is better able to do this than low-frequency training. As weight decay increases, frequency of training tends to become more important in determining eventual error scores, and age of acquisition becomes less important. Thus, when weight decay was absent in Simulation 14 or very small, early low-frequency patterns trained once per epoch had lower error scores after extensive training than late patterns trained 10 times per epoch (Simulations 11.2, 15.1, and 15.2). As weight decay increased, frequency of training became more influential in determining final error scores (Simulations 15.3 and 15.4). This represents only a modulation of age of acquisition and frequency effects, however: Both effects were present and influenced network structure to a significant degree at all levels of weight decay that allow the network to learn all patterns in the training set.

It is possible that by reducing the tendency of a network to become rigid, weight decay may make it more receptive to very late items. Simulation 13 suggests that it would be extremely difficult for an adult to acquire new vocabulary. We have little doubt that acquisition of new vocabulary is harder for adults with

their mature networks than for children with their immature networks, but it is perhaps not quite as hard for an adult to learn new words as Simulation 13 might imply. An element of weight decay should make a mature network more receptive to new information as well as improving its capacity to generalize old learning to new inputs (Plaut, 1997; Plaut et al., 1996). That said, the consequences of reducing the training of early, high-frequency patterns (Simulation 12) would also be expected to be more severe with the addition of weight decay than without, so the improved receptiveness of the mature network could come at a cost to information that is acquired early but not regularly refreshed thereafter.

Finally, the lesioned network showed lower error to early than late acquired items (Simulation 16). This may help to explain the effects of age of acquisition on word retrieval in normal aging (Hodgson & Ellis, 1998) and after brain damage in adulthood (Ellis et al., 1996; Hirsh & Ellis, 1994; Hirsh & Funnell, 1995; Lambon Ralph et al., 1998).

The Relationship Between Age of Acquisition and Frequency Effects

Each time a pattern is presented to the network, it has the opportunity to influence network structure. The more often a pattern is presented, the more influence it has, so patterns trained with high frequency are learned better than patterns trained with low frequency. Final performance on a given pattern (association) is not, however, a simple reflection of the number of times that pattern has been trained. This is because early presentations have a greater impact on final network structure than later presentations. Whereas frequency of training is well established in the literature as a general factor underlying learning in connectionist models using distributed representations, the present work shows that the state of the system at the point when training occurs (age of acquisition) is another. Factors such as the presence and strength of weight decay modulate the relative influences of frequency and age of acquisition on the performance of the network, but both factors exert a significant effect under all circumstances that allowed the network to learn the full set of training patterns. As far as we are aware, no one has previously drawn attention to the fact that age of acquisition effects emerge out of the basic properties of distributed memory networks at least as naturally as frequency effects do.

The fact that training has more influence when the network is young than when it is old means that age of acquisition and frequency effects cannot be combined in terms of simple cumulative frequency (cf. Gilhooly, 1984; Lewis, 1999). That said, the simulations reported here suggest that it may be possible to accommodate age of acquisition and frequency within a single framework. The point of entry of a pattern into training is one factor that affects the ultimate quality of its representation; another is the frequency with which it is trained. Much has been made in the past of the parallel between frequency effects in connectionist networks and in human performance (e.g., Monsell, 1991; Plaut et al., 1996; Seidenberg & McClelland, 1989), even though many of the classic experimental studies of frequency effects were confounded by differences in age of acquisition (Morrison & Ellis, 1995).

Within the type of network used here, differences in the point of entry of patterns into training (age of acquisition) and differences in the frequency with which patterns are subsequently trained affect network structure in essentially the same way: by influencing the extent to which weights change in response to training. Early training results in larger weight changes, as does frequent training. Network structure is determined by point of entry into training and frequency of training, so patterns entered late or trained with low frequency struggle to reconfigure the network in ways that would optimize their representation. Thus, although we have shown that age of acquisition effects cannot be reduced to simple cumulative frequency, a common account may nevertheless be possible for the two effects.

We have shown that the precise nature of the interaction between age of acquisition and frequency is affected by such factors as the differential between high and low frequency (Simulation 11) and the presence and strength of weight decay (Simulation 15). Note, though, that the ways in which age of acquisition and frequency affect the network remain the same in all of these simulations: The presence or absence of a statistical interaction between two variables cannot be taken as providing any clear evidence that the variables in question affect the same or different processing levels or mechanisms, at least not in the case of artificial neural networks. The present analyses also carry the strong implication that any task that is affected by age of acquisition will also be affected by frequency, and vice versa. It has sometimes been claimed that certain tasks are affected by one of these factors and not the other, but subsequent research has tended to support the view that both exert an influence. For example, early studies of object naming by Carroll and White (1973), Gilhooly and Gilhooly (1979), Morrison et al. (1992), and others often reported effects of age of acquisition but not frequency, but more recent studies that have typically involved more items and better measures of word frequency have usually revealed effects of both variables (e.g., Barry et al., 1997; Ellis & Morrison, 1998; Snodgrass & Yuditsky, 1996). Similarly, some studies of word naming have reported effects of age of acquisition but not word frequency (e.g., Brown & Watson, 1987; Morrison & Ellis, 1995), but more recent studies again indicate that both variables exert an effect (Gerhand & Barry, 1998, 1999a, 1999b; Morrison & Ellis, 2000). The account offered here would predict that future studies will show that tasks affected by one factor will also be affected by the other.

Generality of Age of Acquisition Effects

We have demonstrated that age of acquisition effects are a fundamental property of a type of connectionist network that has been widely used to model human cognitive processes and that these effects emerge out of network principles at least as naturally as effects of frequency, imageability, or spelling–sound regularity. This property appears not to have been commented on before or linked to age of acquisition effects in human lexical processing, presumably because simulations have usually involved training networks from the outset on all of the patterns to be learned (or entirely replacing one set of patterns with another in the case of simulations relating to learning and catastrophic interference). Much human learning, however, is cumulative, including human vocabulary acquisition, and if networks are to simulate natural learning, they need to be trained cumulatively.

We used a three-layer network involving back-propagation partly because it is such a widely used species of network and partly because some authors have claimed that age of acquisition effects are incompatible with back-propagation networks (Gerhand & Barry, 1998; Moore & Valentine, 1998; Morrison & Ellis, 1995). We suspect, however, that age of acquisition effects will be found to characterize a wide range of artificial neural networks. Specifically, age of acquisition effects will be found whenever early training has more of an impact on network structure than later training. Under those circumstances, any network will become more rigid and less plastic as learning proceeds, with the consequence that material introduced late into training will be more difficult to learn than material introduced early (cf. Munro, 1986). It may be possible to lessen this tendency in various ways, but we would argue that the strength of the evidence for the existence of age of acquisition effects means that the psychological relevance of any network or learning algorithm that fails to manifest such effects must be called into question.

Self-organizing (Kohonen) networks have attracted growing interest in recent years (e.g., Anderson, 1999; Luckman, Allinson, Ellis, & Flude, 1995; MacWhinney, 1998). A self-organizing network is an unsupervised neural network that uses competitive learning to create a two-dimensional topographic map of the input in which similar input patterns are represented by units that are physically close together on the map and dissimilar input patterns are pushed apart (Kohonen, 1984, 1990; Ritter, 1995). Morrison (1993), Morrison and Ellis (1995), and Ellis and Morrison (1998) suggested that patterns introduced early into the training of self-organizing networks will colonize the entire map, with the result that patterns introduced later will have to be squeezed in around them and will develop less effective representations. Thus, in self-organizing networks, differences in age of acquisition might be shown to affect the representations themselves, not just the ease with which representations at one level can activate associated representations at other levels (as is the case for the simulations reported here). If future research were to support Brown and Watson's (1987) conjecture that the phonological representations of late acquired words are more fragmented than those of early acquired words, then self-organizing networks might offer some insights into how this could come about.

Although age of acquisition effects have largely been documented in the domain of word recognition and production, there is nothing in the account of age of acquisition effects we have offered here to suggest that such effects will be confined to that domain. In fact, our account reveals quite the opposite: Age of acquisition effects should occur whenever networks with certain basic properties are required to learn and represent associations between input and output patterns in a cumulative and interleaved manner. Those associations that are able to influence the structure of a relatively unformed network will have a greater impact on its final structure than associations learned by a network that has already acquired some knowledge and consequent organization.

Munro (1986) suggested that the critical periods seen in a range of developmental domains (for example, in ocular dominance in the visual cortex) might reflect progressive reductions in neural plasticity under conditions of cumulative learning. Age of acquisition effects have recently been reported in situations that do not involve recognizing and producing words. Vitkovitch and Tyrell (1995) reported that age of acquisition of object names predicted speed of responding in a task in which participants had to indicate whether pictured objects were real or unreal (chimeric combinations of two halves of different objects). They suggested that age of acquisition of object names is likely to correlate with the age at which the objects themselves are encountered (so that one learns the early word *sheep* around the time when one first sees a real or pictured sheep, and one learns the word *microscope* around the somewhat later time when one first sees a real or pictured microscope). Moore and Valentine (1999) likewise reported an age of acquisition effect in a task requiring participants to judge faces as familiar (celebrity faces) or unfamiliar. Reaction times were faster for the faces of celebrities that participants rated as having been learned early in life than for the faces of equally familiar celebrities rated as having been learned more recently. Similarly, Lewis (1999) had participants classify characters from two well-known television soap operas according to which program they appeared in and reported an effect of age of acquisition (time since the character entered the soap opera) over and above an effect of frequency of appearance.

If our analysis is correct, then age of acquisition effects should occur whenever learning is cumulative and accompanied by a gradual decline in the plasticity of the network responsible for learning patterns and associations. Factors yet to be investigated may influence the impact that network structure created by early patterns exerts on the assimilation of later patterns. For example, when mappings between input and output patterns are highly consistent, as in the reading of English words with predictable spelling–sound correspondences or almost all words in languages such as Italian or Spanish, age of acquisition effects may be reduced because late acquired words should be able to exploit the network structure generated by early words. In contrast, when late words require new or different input–output connections, as in reading late acquired exception words in English or learning new object names in any language, age of acquisition should be a major factor in determining processing speed and accuracy. The adult human lexical system, like the human object and face recognition systems, is created by a process of gradual, cumulative learning within a neural network. It is our contention that networks trained in this way preserve within their structure the vestiges of their creation, vestiges that reveal themselves in adulthood as age of acquisition effects.

References

Anderson, B. (1999). Kohonen neural networks and language. *Brain and Language, 70,* 86–94.

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency and name agreement. *Quarterly Journal of Experimental Psychology, 50A,* 560–585.

Brown, G. D. A., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition, 15,* 208–216.

Butler, B., & Hains, S. (1979). Individual differences in word recognition latency. *Memory & Cognition, 7,* 68–76.

Carroll, J. B., & White, M. N. (1973). Word frequency and age-of-acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology, 25,* 85–95.

Cuetos, F., Ellis, A. W., & Alvarez, B. (1999). Naming times for the Snodgrass and Vanderwart pictures in Spanish. *Behavior Research Methods, Instruments, and Computers, 31,* 650–658.

Ellis, A. W., Lum, C., & Lambon Ralph, M. A. (1996). On the use of regression techniques for the analysis of single case aphasic data. *Journal of Neurolinguistics, 9,* 165–174.

Ellis, A. W., & Morrison, C. M. (1998). Real age of acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 515–523.

Fahlman, S. (1989). Fast learning variations on back-propagation: An empirical study. In M. C. Mozer, P. Smolensky, D. Touretzky, J. L. Elman, & A. Weigand (Eds.), *Connectionist models: Proceedings of the 1988 Summer School* (pp. 38–51). San Mateo, CA: Morgan Kaufman.

Gerhand, S., & Barry, C. (1998). Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 267–283.

Gerhand, S., & Barry, C. (1999a). Age of acquisition and frequency effects in speeded word naming. *Cognition, 73,* B27–B36.

Gerhand, S., & Barry, C. (1999b). Age of acquisition, frequency and the role of phonology in the lexical decision task. *Memory & Cognition, 27,* 592–602.

Gilhooly, K. J. (1984). Word age-of-acquisition and residence time in lexical memory as factors in word naming. *Current Psychological Research and Reviews, 3,* 24–31.

Gilhooly, K. J., & Gilhooly, M. L. (1979). Age-of-acquisition effects in lexical and episodic memory tasks. *Memory & Cognition, 7,* 214–223.

Gilhooly, K. J., & Watson, F. L. (1981). Word age-of-acquisition effects: A review. *Current Psychological Research, 1,* 269–286.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence, 40,* 185–234.

Hirsh, K. W., & Ellis, A. W. (1994). Age of acquisition and aphasia: A case study. *Cognitive Neuropsychology, 11,* 435–458.

Hirsh, K. W., & Funnell, E. (1995). Those old, familiar things: Age of acquisition, familiarity and lexical access in progressive aphasia. *Journal of Neurolinguistics, 9,* 23–32.

Hodgson, C., & Ellis, A. W. (1998). Last in, first to go: Age of acquisition and naming in the elderly. *Brain & Language, 64,* 146–163.

Kohonen, T. (1984). *Self organization and associative memory.* Berlin: Springer-Verlag.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78,* 1464–1480.

Lambon Ralph, M. A., Graham, K. S., Ellis, A. W., & Hodges, J. R. (1998). Naming in semantic dementia—What matters? *Neuropsychologia, 36,* 775–784.

Lewandowsky, S. (1991). Gradual unlearning and catastrophic interference: A comparison of distributed architectures. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honour of Bennet B. Murdock* (pp. 445–476). Hillsdale, NJ: Erlbaum.

Lewis, M. B. (1999). Age of acquisition in face categorisation: Is there an instance-based account? *Cognition, 71,* B23–B39.

Luckman, A. J., Allinson, N. M., Ellis, A. W., & Flude, B. M. (1995). Familiar face recognition: A comparative study of a connectionist model and human performance. *Neurocomputing, 7,* 3–27.

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology, 49,* 199–227.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the success and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises.* Boston: MIT Press.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109–164). New York: Academic Press.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 148–197). Hillsdale, NJ: Erlbaum.

Moore, V., & Valentine, T. (1998). The effect of age of acquisition on speed and accuracy of naming famous faces. *Quarterly Journal of Experimental Psychology, 51A,* 485–513.

Moore, V., & Valentine, T. (1999). The effects of age of acquisition in processing famous faces and names: Exploring the locus and proposing a mechanism. In *Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society, Vancouver, 1999* (pp. 416–420). Mahwah, NJ: Erlbaum.

Morrison, C. M. (1993). *Loci and roles of word age of acquisition and word frequency in lexical processing.* Unpublished doctoral dissertation, University of York, York, England.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology, 50A,* 528–559.

Morrison, C. M., & Ellis, A. W. (1995). The roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 116–133.

Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology, 91,* 167–180.

Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition, 20,* 705–714.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165–178.

Munro, P. W. (1986). State-dependent factors influencing neural plasticity: A partial account of the critical period. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models* (pp. 471–502). Cambridge, MA: MIT Press.

Nagy, W. E., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24,* 262–282.

Nickels, L., & Howard, D. (1995). Aphasic naming: What matters? *Neuropsychologia, 33,* 1281–1303.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes, 12,* 765–805.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377–500.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285–308.

Ritter, H. (1995). Self-organising feature maps: Kohonen maps. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 846–851). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning

internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Sharkey, N. E., & Sharkey, A. J. C. (1995). An analysis of catastrophic interference. *Connection Science, 7,* 301–329.

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments, and Computers, 28,* 516–536.

Turner, J. E., Valentine, T., & Ellis, A. W. (1998). Age of acquisition, not word frequency, affects auditory lexical decision. *Memory & Cognition, 26,* 1282–1291.

Veitch, A. C., & Holmes, G. (1991). A modified Quickprop algorithm. *Neural Computation, 3,* 310–322.

Vitkovitch, M., & Tyrell, L. (1995). Sources of disagreement in object naming. *Quarterly Journal of Experimental Psychology, 21A,* 1155–1168.

Received January 13, 1999
Revision received March 1, 2000
Accepted March 2, 2000 ■