In: Keating, J.G. (Ed.), 1995, *Neural Computing Research and Applications III (Proceedings of the Fifth Irish Neural Networks Conference),* 62-69. Maynooth, Ireland: St. Patrick's College.

Modelling Lexical Decision: Who needs a lexicon?

John A. Bullinaria

Centre for Speech and Language, Department of Psychology
Birkbeck College, Malet Street, London WC1E 7HX

j.bullinaria@psyc.bbk.ac.uk

Abstract

The problem of modelling lexical decision in connectionist models is discussed. It is shown how lexical decision can be performed by a simple neural network with no explicit lexicon and no recurrent connections. We also see how simulated reaction times can be extracted from such systems that are in broad agreement with various experimental data concerning semantic and associative priming.

1. Introduction

The basic language processing tasks of reading and spelling have become important benchmark tests of connectionist cognitive modelling and also of connectionism more generally. Already connectionist models perform well at these tasks in terms of learning the training data and in generalizing to new items. We now need to test and constrain these models further by comparing them with human performance on the various experimental conditions that psychologists have designed to probe the representations and processing procedures employed by the human brain. Of particular interest are the results of reaction time studies of naming (i.e. producing a pronunciation) and lexical decision (i.e. deciding whether a string of letters or phonemes is a real word).

It is well known that humans are able to perform both naming and lexical decision quickly and with high accuracy and the corresponding experiments (which investigate errors, reaction times and priming) place particularly strong constraints on connectionist (and other) models. The connectionist modelling of the naming studies was discussed in Bullinaria (1995). In this paper we shall discuss the modelling of the lexical decision studies.

One advantage of connectionist systems over others (such as symbolic analogy models) is their ability to generalize well to novel items, i.e. to perform nearly as well with novel inputs as with those items on which they were trained. For example, any successful network model of reading must (like humans) be able to read aloud words or non-words it has never seen before. However, if the model can deal with new words in the same way as those it was trained on, it will be difficult for that system to perform lexical decision without the introduction of an explicit lexicon. It was for this reason that the earliest single route connectionist models of reading aloud (Sejnowski & Rosenberg, 1987; Seidenberg & McClelland, 1989) were unable to perform reliable lexical decision (Besner, Twilley, McCann & Seergobin, 1990). There was simply no criteria within those systems that could form the basis for lexical decision. The same is true of the more recent and more successful direct route models of reading and spelling (Bullinaria, 1994a; Plaut et al., 1994; Bullinaria, 1994b).

It is now generally agreed (e.g. Coltheart et al., 1993; Bullinaria, 1994b; Plaut et al., 1994) that such single route models must be supplemented by some form of additional semantic or lexical route so that we end up with what is often described as a Dual Route Model (e.g. Coltheart et al., 1993). Some preliminary work has been carried out on the semantic/lexical route of such models (e.g. Plaut & Shallice, 1993; Plaut et al., 1994), but it is still far from clear if an explicit lexicon is required or if some more distributed semantic system is sufficient (e.g. Masson, 1991).

In the remainder of this paper we shall investigate some simple feed-forward networks that map between orthography and phonology and semantics and show that under certain (but not all) circumstances it *is* possible for connectionist systems to perform reliable lexical decision without an explicit lexicon. It will also be shown how these systems can produce reaction time and priming results that are in broad agreement with experiments on humans. This work provides the foundations for more detailed morphological modelling currently being carried out (Bullinaria & Marslen-Wilson, in preparation) and the construction of complete connectionist dual route models of reading and spelling (Bullinaria, in preparation).

2. Modelling Lexical Decision

Within the connectionist framework, there are currently two main approaches to modelling the lexical/semantic system. The original approach uses a localist representation with a single node for each word (e.g. McClelland & Rumelhart, 1981). Lexical decision is then simply a matter of checking to see if, and how quickly, any of the word nodes are activated. Priming (i.e. the effect by which response is speeded by prior presentation of certain related words) can arise due to activation passing between appropriately related nodes. More recently, Hinton & Shallice (1991) and Plaut & Shallice (1993) have shown how a reading model using distributed semantic representations can provide an account of deep dyslexia. In this distributed approach, priming may (depending on the details of the model) still arise due to activation passing between the nodes which now correspond to semantic microfeatures. However, it is now also possible for priming to occur simply due to the overlap of semantic representations that will inevitably occur for semantically related words.

In many ways the distributed approach seems more natural in the connectionist framework (see Plaut,

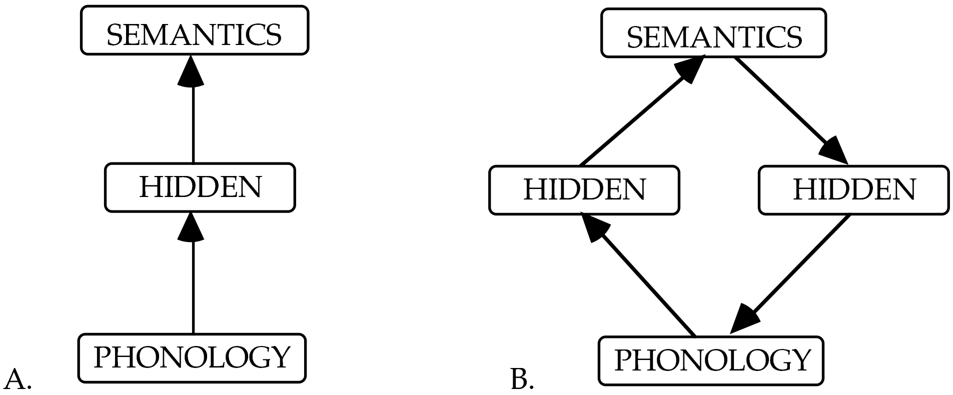

Figure 1. Simple feed-forward networks for mapping phonology to semantics.

1995, for further discussion), but performing the lexical decision becomes a much less straightforward matter. If our system still involved some form of explicit lexicon, the lexical decision could be carried out by checking the target word or the corresponding semantic activation against this lexicon. If the lexicon search were ordered by frequency in some way, then we would automatically arrive at the experimental result of faster decision times for high frequency words than for low frequency words than for non-words (Forster, 1976). However, there is considerable experimental evidence against such sequential search models (e.g. Marslen-Wilson, 1987). There is still the possibility of a realistic non-sequential lexical search within a connectionist system, but we shall see if we can manage without such a search altogether.

If the whole lexicon is replaced by a distributed system of semantic microfeatures, we need some other method of performing lexical decision. It is natural to consider the reaction times to be the time taken for the relevant sets of microfeatures to be activated or to settle down to or become sufficiently close to their stable state. The awkward question of determining whether this final state corresponds to a word or a non-word is rarely investigated in detail. One could argue that there is some form of (possibly variable) processing deadline, such that failure to activate or stabilise by that time indicates a non-word (e.g. Coltheart et al., 1977). This would certainly explain the faster response for words over non-words, but it is not obvious that is it possible to find a deadline that accurately separates the two classes. Similarly, we could consider lexical decision to be based on simple familiarity, and argue that the networks output activation error scores are a reasonable measure of this familiarity. A potential problem with using such error scores is that, if there is no explicit lexicon to check against, it is difficult to see how the error scores could be calculated.

To test the feasibility of these various possibilities we really need to explicitly implement and test some representative models. The simplest model we could consider is essentially the same for the phonology to semantics mapping as for orthography to semantics: A static input coding of phonology/orthography is associated by gradient descent learning with a static output semantic vector in a simple feed-forward network.

For mono-syllabic words the relationship between phonology (or orthography) and semantics is essentially random. Some progress has been made towards generating realistic semantic representations (e.g. Schütze, 1993; Lund, Burgess & Atchley, 1995), but for simplicity here we shall use conveniently constructed random semantic vectors. Each of these vectors is taken to be 27 dimensional binary with exactly three bits on. These three bits are supposed to correspond to the activated semantic microfeatures of Hinton & Shallice (1991) and Plaut & Shallice (1993). For the phonology we shall take one unit for each of the 20 most common onset consonant clusters, the 10 most common vowel clusters and the 20 most common offset consonant clusters. This gives us a simplified version of the phonological representation used by Plaut et al. (1995). This kind of representation might not be sufficient to give human level reading and spelling performance and cannot even deal with all English mono-syllables, but it will be sufficiently representative for our purposes here. We shall use a fully connected feed-forward network (as shown in Figure 1A) with 400 hidden units and train it on a random set of 200 mono-syllables taken from the Seidenberg & McClelland (1989) corpus. The training procedure was standard back-propagation with cross-entropy error function, a learning rate of 0.005 and no momentum (Hinton, 1989).

The network was trained for 5000 epochs with all 200 training words appearing in random order in each epoch. By epoch 2500 all the training vectors had been learnt correctly, i.e. the three on bits each had activation greater than 0.5 whilst all the off bits had activation less than 0.5. The remainder of the training reduced the output activation error scores further. The error score distributions at each stage were the usual noisy positively skewed gaussians. The cross-entropy error scores were highly correlated (Pearson r = 0.95) with the standard root mean squared errors (RMSE) so we shall only discuss the RMSE in the following.

Our main purpose here is to establish if the network's outputs for non-words are sufficiently and consistently different in some way from those for the training words that they may be used as the basis for lexical decision. Because of our reduced phonological representation, there are only 4000 possible input strings for our network. We took 200 'words' for training, leaving us 3800 'non-words' to use for testing. We shall look at the network at both 2500 epochs and 5000 epochs to give us an idea of how things change as a result of more or less training. In more realistic training data with a wide word frequency distribution many of the low frequency words will be more poorly learnt as at our epoch 2500. The higher frequency words will be better learnt as at our epoch 5000.

The first thing to consider is the total semantic activations produced by each input. For our training words the sum squared activation will obviously be close to 3. At epoch 2500 we actually find 2.58 ± 0.15 and by epoch 5000 this has increased to 2.93 ± 0.03. If the semantic activations for non-words were much less than this, we might have a basis for lexical decision. Unfortunately they are not – we find 2.42 ± 1.07 at epoch 2500 and 2.53 ± 1.13 at epoch 5000. If we also take into account the fact that realistic semantic representations are unlikely to have equal numbers of activated micro-features for every word, we see that this approach is not really viable.

Although there is a very large overlap between the total semantic activations for the words and non-words, the words have been learnt to be near-binary, whereas the non-words have a wider distribution, so maybe this could be used for lexical decision At epoch 5000 the sum squared deviation from binary is 0.00096 ± 0.00052 for the words compared with 0.23 ± 0.17 for the non-words and there are only 43/3800 (1.1%) non-words overlapping with the word distribution. Unfortunately at epoch 2500, things don’t look so good. The deviation from binary is then 0.036 ± 0.021 for the words and 0.31 ± 0.19 for the non-words with 568/3800 (14.9%) overlap. It was a similar overlap that also made it impossible for the Seidenberg & McClelland (1989) model to perform reliable lexical decision (Besner et al., 1990).

Another problem is that in realistic semantic representations there is no good reason to suppose that we will have binary activations for each word, i.e. each unit either fully on or fully off. Shades of meaning and variable contexts suggest that we cannot necessarily assume binary activations of the microfeatures. Certain microfeatures may well be activated to intermediate degrees. This will clearly make it difficult to perform lexical decision based purely on semantic activations without some form of lexicon to check the activations against.

We know that it is possible for non-words (e.g. ‘slithey’) to elicit semantic activation, particularly if they are closely related orthographically or phonologically to real words. It is also natural that, when we hear a non-word mixed in with normal speech, we should interpret it semantically as best we can and if necessary allow the available context information to adjust its meaning so that the sentence makes sense. If we hear a slightly mis-pronounced or slightly garbled word we do not reject it completely and fail to parse the sentence. Indeed, we often fail to detect such errors at all (Cole, 1973). It should be considered good to see our network behaving in a similar manner. However, even though context free lexical decision is a somewhat artificial task, we still need to understand how humans can do it so well.

We have seen that simple phonology to semantics networks on their own have a hard time trying to perform lexical decision. However, one possible alternative procedure that could be implemented in these simple models is to consider the internal consistency of the activations. For example, if a phonological input produces a particular pattern of semantic activation, we could check that this pattern of activation would in turn produce the same phonological activation. It is natural in the cascaded approach for such activation to develop automatically even though it is not necessarily going to be used for anything. To implement this we need to train the network separately to map from phonology to semantics as well as from phonology to semantics (as in Figure 1B). We can then feed the activation through the four stages from phonology back to phonology and check for consistency. This is similar to the use of orthographical error score for lexical decision in the Seidenberg & McClelland (1989) reading model. If the training data corresponded to a fairly regular mapping, the consistency for non-words would be a measure of the generalization ability of the system and we would expect it to be high. (This is why the procedure didn’t work very well in the Seidenberg & McClelland model.) However, for an essentially random mapping such as phonology to semantics, we can expect the consistency to be low for non-words; possibly low enough to perform lexical decision.

Training our extended network with the same training data and parameters as above leads to an accurate phonology to semantics to phonology mapping and we can examine the phonological output activations in the same way as the semantic outputs above. The difference is that we now have the input phonology for comparison.

Not surprisingly there were no consistency errors (i.e. false negatives) for the training words at either epoch 2500 or epoch 5000. The number of consistency errors (i.e. false positives) on the non-words at epoch 5000 was 28/3800 (0.74%) in terms of best activated phonology. If we look at the phonological output error scores, the overlap between words and non-words is 0/3800 (0.00%). At epoch 2500 the number of consistency errors is still only 28/3800 (0.74%) and the overlap of error scores is 6/3800 (0.16%). If we train the network with 400 random words instead of 200, the number of consistency errors falls to 8/3600 (0.22%) with 0/3600 (0.00%) overlap of error scores even for the epoch 2500 case. It would appear that this consistency checking procedure does provide a fairly reliable method of performing lexical decision without an explicit lexicon.

We next need to consider how lexical decision reaction times can be extracted from our network models and how we can simulate priming experiments.

3. Modelling Reaction Times

In the earlier models of reading (Seidenberg & McClelland, 1989) it was argued that reaction times (i.e. naming latencies) could be simulated by the networks output activation error scores. This indeed produced many of the frequency, regularity and consistency effects exhibited by adult readers. More recently it has been argued (Bullinaria, 1995) that a more principled account of reaction times can be provided by considering the time it takes for activation to cascade through a multi-layer network (McClelland, 1979). This approach also has the advantage that it can provide direct accounts of priming and speed-accuracy trade-off effects (Bullinaria, 1995).

Within the cascaded approach the natural system of equations to describe the build-up of activation is:

$$Out_i(t) = Sigmoid(Sum_i(t))$$

$$Sum_i(t) = Sum_i(t-1) + \sum_j w_{ij} Prev_j(t) - \lambda Sum_i(t-1)$$

so that at each discrete time slice t the output $Out_i(t)$ of unit i is the usual sigmoid of the sum of the inputs into that unit at that time. The sum of inputs $Sum_i(t)$ is given by the existing sum at time t–1 plus the additional weight w_{ij} dependent contribution fed through from the activation $Prev_j(t)$ of the previous layer and a natural exponential decay of activation depending on some decay constant λ. These are equivalent to the equations commonly used to update the state of recurrent networks (e.g. Plaut, 1995). In the asymptotic state $Sum_i(t) = Sum_i(t-1)$, so we have:

$$Sum_i(t) = \sum_j \frac{w_{ij}}{\lambda} Prev_j(t) .$$

It follows that the asymptotic state of our cascaded network is equivalent to a standard feedforward network with weights w/λ. Thus, assuming the right way to train the cascading network is to adjust the weights so that it produces the right asymptotic output for each input, we can obtain exactly the same results by training the standard feedforward network in the conventional manner, e.g. by back-propagation. In this way, any back-propagation network can be trivially re-interpreted as a cascaded network and we can easily extract reaction times from it by counting the time steps required for the outputs to reach a given threshold after a particular pattern of activation is presented at the inputs.

We can model speed-accuracy trade-offs by adjusting these thresholds and model various forms of priming by allowing the network activations to update smoothly after a change of inputs according to the above equations. Both of these effects are difficult to simulate in conventional feedforward networks.

4. Modelling Lexical Decision Reaction Times

We can now apply the cascading activation equations to the above feed-forward networks and extract the lexical decision reaction times (RTs). All our RTs and other time durations are simply the number of time slices times the time scale parameter λ which we shall always take to be 0.1.

First it is instructive to consider how the output activations vary as one input word is replaced by another in the simple phonology to semantics mapping of Figure 1A. Figure 2 shows how *Sum(t)* varies for some representative semantic units after the word /keg/ is replaced by the word /sok/. We see that the behaviour is rather complex and it is not obvious what the appropriate measure of RT should be. We could follow Plaut (1995) and take the RT to be the time taken for the network to settle to within some tolerance of the stable state, but our RTs could then be delayed by units that actually have little influence on the lexical decision process. Instead, we could follow Bullinaria (1995) and take the time for the integrated output activations to reach some threshold, but this makes less sense if we allow the possibility of non-binary semantic vectors. Fortunately, by considering above how our networks can actually perform lexical decision, we have also provided the answer to the question of how to extract the RTs. The RT for words is simply the time taken for the difference between the output and input phonology to reduce to an appropriate threshold. Non-words, of course, should never reach this threshold and their finite RTs will presumably be due to some processing deadline (Coltheart et al., 1977). Even if this time limit is variable and/or augmented by explicit mismatch detection, it is easy to account for the longer RTs for non-words than words.

We tested our networks using a conservative phonological error threshold of 0.2 and a generous time cut-off of 10 time units to allow for words that have only just been learnt. At both epoch 2500 and 5000, all words were given the correct lexical decision response. For the non-words, all but 16/3800 (0.4%) were given a negative response at epoch 2500 and all but 22/3800 (0.6%) at epoch 5000. The RT distributions for the words were realistic noisy skewed gaussians with mean 6.26 (s.d. 0.77) at epoch 2500 and 4.84 (s.d. 0.23) at epoch 5000. The reduction of RTs with training is consistent with experimental observations of faster RTs for higher frequency words.

We also checked how well these consistency RTs correlated with other simulated RTs, namely output error scores (Seidenberg & McClelland, 1989; Moss et al., 1994), settling times (Plaut, 1995) and integrated activation build-up times (Bullinaria, 1995). At epoch 2500 the respective correlations with the consistency RTs are r = 0.92, 0.90 & 0.95 and at epoch 5000 we have r = 0.58, 0.87, 0.95. Thus using error scores rather than a full cascaded approach begins to look slightly suspect at later stages of training. Another important question is how do the RTs at the semantic layer correlate with our complete lexical decision RTs? These correlations are quite low (r = 0.45, 0.47 & 0.46 at epoch 2500 and r = 0.28, 0.36 & 0.32 at epoch 5000) suggesting that it might not be wise to ignore the details on what goes on beyond the semantic activations when simulating lexical decision.

Finally, we must check that the RTs at the semantic layer cannot be used for lexical decision in a way that the activations themselves could not. If we use the maximum settling time found for the words as the reaction deadline, we can count the number of non-words that settle before this as the number of false positive decisions. This turns out to be 150/3800 (3.9%) at epoch 5000 and 2877/3800 (75.7%) at epoch 2500. We can reduce these error rates slightly by reducing the deadline and allowing false-negatives, but it is clear that this is not a reliable way of performing lexical decision.

5. Modelling Semantic and Associative Priming

From an experimental point of view, there is still considerable disagreement in the literature concerning the precise nature of associative and semantic priming; for example, compare Shelton & Martin (1992), Moss

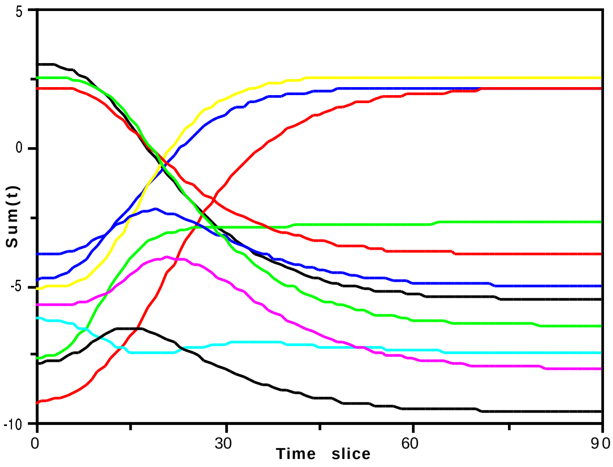

Figure 2. Time course of typical semantic activations.

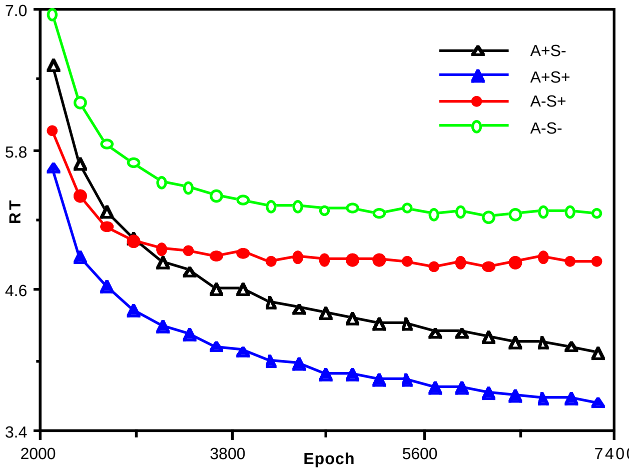

Figure 3. Graph of RTs during training.

et al. (1995) and Lund et al. (1995). For this reason we shall simply assume that both types of priming exist and outline the performance of our model without a detailed comparison with the conflicting experimental results.

As noted above, priming by semantically related words will arise naturally in our model due to the overlap of semantic features. To test this explicitly, our 200 training words were split into 40 sets of five phonologically unrelated words. Each set of five words consisted of one target word, two semantically related words (with two of the three on bits in common with the target) and two semantically unrelated words (with no on bits in common with the target). At epoch 2500 the mean target RTs was 5.46 (s.d. 0.99) when preceded by semantically related primes compared with 6.54 (s.d. 1.03) for the unrelated control primes. Similarly at epoch 5000 we had mean RTs of 3.66 (s.d. 0.55) and 4.64 (s.d. 0.47). The facilitation by the semantically related primes was highly significant ($p < 0.0001$) in both cases. (In this, and all the following priming simulations, there were no lexical decision errors and significance was measured using standard t-tests.)

Priming is also found to be produced by words that are associated but not semantically related (e.g. ‘pillar’ primes ‘society’). One approach that has been used to model such associative priming in the past (Moss et al., 1994; Plaut, 1995) relies on the order in which the words appear during the training process. If the word ‘dog’ immediately follows the word ‘hot’ during training much more often than it should by chance, then it can be expected that any efficient learning system will come to process the word ‘dog’ more quickly when following ‘hot’ than when following some other word. That is, it will exhibit associative priming. (We shall follow Moss et al. and Plaut in not worrying, at this stage, about what happens to the intervening words that sometimes come between associated pairs, such as the ‘of’ in ‘pillar of society’.)

Both Moss et al. (1994) and Plaut (1995) have recurrent connections in their networks that allow them to ‘remember’ such useful between words information and hence exhibit this effect. Clearly, our simple feed-forward networks cannot operate in this way. However, if we take the cascaded activation approach more seriously, we can see how associative priming might occur. First, our phonology to semantics network should be considered as just one stage of a multi-stage system from the basic sound detectors in our ears right through to the motor control system that drives the lexical decision response. It thus makes sense for our phonological activations to build up and decay over a period of time rather than instantaneously appearing and disappearing. It also makes sense for the learning process to be carrying on throughout the cascading process rather than just at the endpoints. The asymptotic states that the network is learning will be the same, but the time courses may be different and the network will have the opportunity to take advantage of associated words to improve its overall performance. In terms of Figure 2, the network should be able to learn to cross-over the crucial semantic activations more quickly and hence exhibit associative priming.

To test this, we used the same 200 training words split into 40 sets of five as before. In each set of five words, one of the semantically related words and one of the semantically unrelated words were chosen at random to be associates of the target word. The target words were preceded by one of their associates on half of their training presentations. Thus each associate was followed by its associated target 25% of the time compared with 0.5% of the time for all the other words. The phonological input activations were allowed to build up linearly over 1.5 time units. Each word was trained until it reached an integrated output activation threshold of 50.0 or a maximum of 15.0 time units. The word’s input activation then decayed over 1.5 time units whilst the next word’s input activation built up. There was no re-setting of hidden unit or output activations between words – all changes came about according to the above cascading equations. Since each word was trained for many time slices (of the order of 100 at a time) the learning rate was reduced to 0.00005. All the other network parameters remained the same.

Figure 3 shows how the mean RTs for our target words, with the four different prime (i.e. preceding word) types, change during training. (A± means associate/non-associate, S± means semantically related/unrelated.) We see that the control primes (A–S–) and purely semantically related primes (A–S+) give target RTs that soon level off, whilst the associated primes (A+S– and A+S+) result in target RTs that continue to reduce as training continues. The amount of priming in

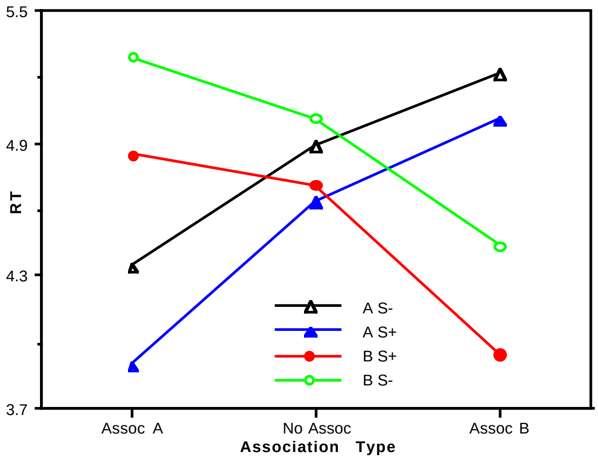

Figure 4. Associative priming control conditions.

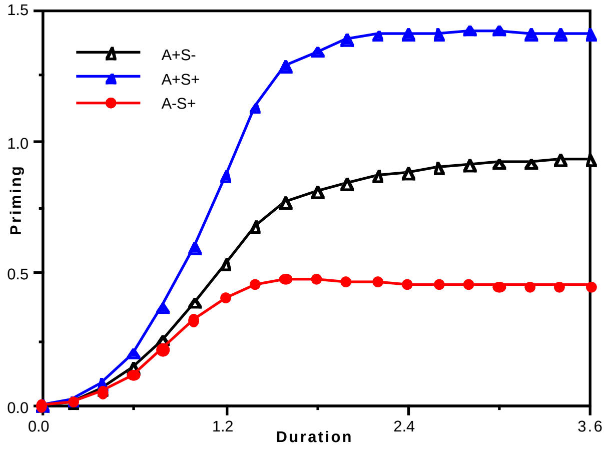

Figure 5. Effect of prime duration on priming.

each case is simply the reduction in RT compared to the RT with the control prime.

It is clearly important that we run appropriate control conditions, especially for small training and testing sets such as ours. The obvious control is to simply repeat the training without the associations. However, it actually proved rather difficult to split the words into two sets such that they were matched throughout training. For this reason, a third run was carried out with the original control set B associated and the original associated set A non-associated. The resulting mean RTs at epoch 5000 for the three runs are shown in Figure 4. This suggests that the associative priming arises at the expense of the reaction times with non-associated primes, though this may simply be an artefact of the number and strength of the associations in our training data. We should not, at this stage, attach too much importance to the precise values of the various primings. The associative priming clearly depends on the degree of association we use and also on the amount of training. The semantic priming depends on the degree of overlap of our rather artificial semantic vectors. However, these results show quite clearly that our model can exhibit a pattern of semantic priming, associative priming and the ‘associative boost’ for mixed priming, qualitatively in line with experiment (Moss et al., 1994; Moss et al., 1995). In the associated conditions the four sets of RTs are all highly significantly different from each other ($p < 0.0001$) except for the unassociated semantic difference in the B set associated run (where $p = 0.005$). In the non-associated condition we have significant semantic priming ($p < 0.001$) with association set differences at most marginal ($p = 0.05$ for S–, $p = 0.3$ for S+).

One worry about performing the lexical decision at the level of phonology is that all our priming results will be swamped by unrealistically large phonological priming. To test this we took 125 words from our training set as targets and for each used two phonologically unrelated primes as controls and two phonologically related primes (with no semantic or associative relations). These showed highly significant phonological priming ($p < 0.0001$), but the effect was small (0.22) compared with our semantic (0.43), associative (0.94) and associative + semantic (1.39) priming.

Following Plaut (1995), we now investigate four important properties of our model, namely the effect of prime duration, the effect of target degradation, priming spanning an unrelated item and mediated priming.

To test the effect of prime duration we first allow the network to settle into a stable state with no input, then present the prime for varying lengths of time before presenting the target and measuring the RT. Figure 5 shows that all three types of priming initially increase with prime duration and eventually level off as the prime has time to settle into its asymptotic state. Note that this behaviour is rather different to that of attractor networks (Plaut, 1995), where the semantic priming reaches a peak and decays away before the associative priming begins to level off. Unfortunately, the complications that tend to arise, due to strategic effects, in experiments with long prime durations may make experimental tests of this difference rather difficult (e.g. Shelton & Martin, 1992).

In visual priming studies (Neely, 1991; Besner & Smith, 1992), larger priming is found if the targets are degraded (e.g. by reducing the contrast). In our priming model we have assumed, for simplicity, that the complete phonology for each word builds up linearly over time. Given this assumption, together with our random semantic vectors and the restricted phonological coding, our model can be equally well considered as representing a mapping between orthography and semantics as between phonology and semantics. It is thus appropriate to expect primarily visual effects, such as target degradation, to be exhibited by our model. We can easily simulate this effect by assuming that degradation of the target simply slows down the build up of the input activation. Not surprisingly, this produces significant increases in the RTs for all prime types. Figure 6 shows how this interacts with the amount of priming, in broad agreement with experiment.

Next we explore the possibility of priming persisting across an intervening item in our model as has been found in some experiments (e.g. Joordens & Besner, 1992; McNamara, 1992). We carried out the priming simulations exactly as before except that an intervening item was presented between the prime and target. Figure 7 shows the mean amount of priming as a function of duration of the intervening item with

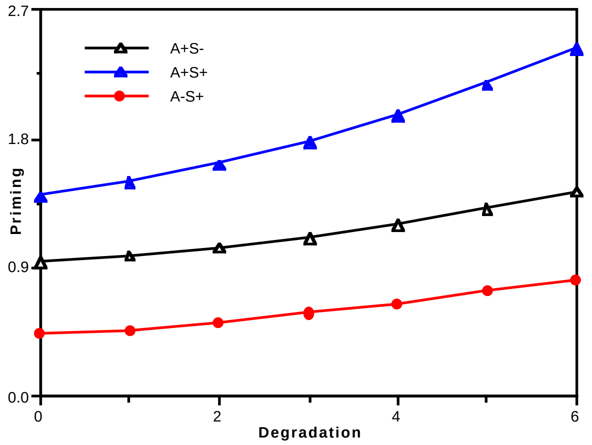

Figure 6. Effect of target degradation on priming.

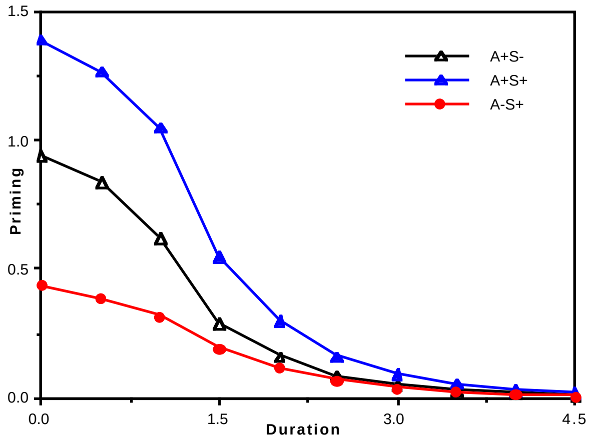

Figure 7. Priming spanning an unrelated item.

averages taken over ten different (phonologically, associatively and semantically unrelated) intervening items in each case. Not surprisingly the amount of priming is seriously reduced by the intervening item, but the priming actually remains highly significant right up to and beyond 4.5 time units, by which time the size of the effect would be far too small to detect under normal experimental conditions. We get a similar reduction in priming (though with slower fall-off) by having a simple (no input) delay between the prime and target.

Finally, we come to the question of mediated priming (e.g. McNamara & Altarriba, 1992; Shelton & Martin, 1992). Although there is no mechanism within our model to account for mediated associative priming (e.g. ‘cordless’ to ‘number’ via ‘phone’), it has been suggested (Matt Davis, personal communication) that it might be possible to have mediated priming involving first association and then semantics (e.g. ‘hot’ to ‘fox’ via ‘dog’). This could arise because a prime that facilitates one particular pattern of semantic activation is also likely to facilitate a closely related pattern of semantic activation. To test this we took the old A+S– primes to be our new primes, so the old A–S+ primes could become our new targets with mediation through our old targets. As controls we took old A+S– primes that were semantically unrelated to the new primes and targets and whose associate was also semantically unrelated to the new target. As before we ensured that there was no phonological overlap within each set of new primes, controls and targets. The result was that small (0.15 time units) but significant ($p < 0.005$) mediated priming was obtained in this way.

6. Conclusions and Discussion

We have described a simple feedforward neural network model that can perform reliable lexical decision without an explicit lexicon. It also allows two distinct causes of priming: semantic priming due to semantic vector overlap and associative priming due to word co-occurrence during learning. The model compares favourably with that of Moss et al. (1994) which showed associative priming but did not exhibit semantic priming and the attractor network model of Plaut (1995) which showed both semantic and associative priming but was not designed to exhibit the associative boost.

There is insufficient space to present the detailed results here, but we find that the build up of semantic activations in the phonology to semantics mapping alone also shows significant semantic and associative priming, which is likely to exhibit itself in any other reliable implementation of lexical decision different to the one discussed here.

One could rightly argue that our parallel phonemic inputs do not take into account the importance of the sequential nature of normal speech input (e.g. Marslen-Wilson, 1987). A variation of the current approach, that does account for many aspects of the time course of semantic activation, involves having the build up of activations in each of the three input blocks occur sequentially - first the onset consonant cluster, then the vowel cluster and finally the offset consonant cluster. Such a modified model still results in significant semantic and associative priming, but to a slightly lesser extent than discussed above. This work will be discussed in more detail elsewhere.

Another limitation of our model is that we have not yet properly explored the important effect of word frequency. These effects come out correctly in Plaut’s (1995) model, but that is no guarantee that they will in ours. Our model needs to be scaled up to more realistic word sets, that include real word frequencies, to check this. We also need to re-run the simulations with more realistic semantic vectors to investigate the effect of having different types and degrees of semantic relation (Moss et al., 1995).

Perhaps the biggest problem left for future research is that of homophones and homographs. Simple phonology to semantics mappings such as ours tend to produce semantic blends for homophonic inputs which cause problems for any form of lexical decision. It is possible that the incorporation of context information or more complex semantic representations will help. In this paper we have simply avoided the problem by using non-homophonic training data, but this problem cannot be avoided for ever.

Acknowledgements

The early stages of this work were carried out while the author was supported by the UK JCI at the Department of Psychology, University of Edinburgh. The current research is supported by the UK ESRC. This work benefited from discussions with Matt Davis, Gareth Gaskell, William Marslen-Wilson, Helen Moss and Lorraine Tyler.

References

Besner, D. & Smith, M.C. (1992). Models of Visual Word Recognition: When Obscuring the Stimulus Yields a Clearer View. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **18**, 468-482.

Besner, D., Twilley, L., McCann, R.S. & Seergobin, K. (1990). On the Association Between Connectionism and Data: Are a Few Words Necessary? *Psychological Review*, **97**, 432-446.

Bullinaria, J.A. (1994a). Connectionist Modelling of Spelling. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 78-83. Hillsdale, NJ: Erlbaum.

Bullinaria, J.A. (1994b) Representation, Learning, Generalization and Damage in Neural Network Models of Reading Aloud. Submitted.

Bullinaria, J.A. (1995). Modelling Reaction Times. In L.S. Smith & P.J.B. Hancock (Eds), *Neural Computation and Psychology (Proceedings of the Third Neural Computation and Psychology Workshop*), 34-48. New York: Springer Verlag.

Cole, R.A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics*, **13,** 153-156.

Coltheart, M., Curtis, B., Atkins, P. & Haller, M. (1993). Models of Reading Aloud: Dual-Route and Parallel-Distributed-Processing Approaches, *Psychological Review*, **100**, 589-608.

Coltheart, M., Davelaar, E., Jonasson, J. & Besner, D. (1977). Access to the Internal Lexicon. In S. Dornic (Ed.), *Attention and Performance VI,* 535-555. Hillsdale NJ: Erlbaum.

Forster, K.I. (1976). Accessing the Mental Lexicon. In R.J. Wales & E. Walker (Eds), *New approaches to language mechanisms*, 257-288. New York: North Holland.

Hinton, G.E. (1989). Connectionist Learning Procedures, *Artificial Intelligence*, **40**, 185-234.

Hinton, G.E. & Shallice, T. (1991), Lesioning an Attractor Network: Investigations of Acquired Dyslexia. *Psychological Review*, **98**, 74-95.

Joordens, S. & Besner, D. (1992). Priming Effects That Span an Intervening Unrelated Word: Implications for Models of Memory Representation and Retrieval, *Journal of Experimental Psychology: Learning, Memory and Cognition,* **18**, 483-491.

Lund, K., Burgess, C. & Atchley, R.A. (1995). Semantic and Associative Priming in High-Dimensional Semantic Space. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, 660-665. Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition*, **25**, 71-102.

Masson, M.E.J. (1991). A distributed memory model of context effects in word identification. In D. Besner & G. Humphreys (Eds), *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Erlbaum.

McClelland, J.L. (1979). On the time relations of mental processes: An examination of systems of processing in cascade. *Psychological Review*, **86**, 287-330.

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.

McNamara, T.P. (1992). Theories of Priming: I. Associative Distance and Lag. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **18**, 1173-1190.

McNamara, T.P. & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. *Journal of Memory and Language*, **27**, 545-559.

Moss, H.E., Hare, M.L., Day, P. & Tyler, L.K. (1994). A Distributed Memory Model of the Associative Boost in Semantic Priming. *Connection Science*, **6**, 413-427.

Moss, H.E., Ostrin, R.K., Tyler, L.K. & Marslen-Wilson, W.D. (1995). Accessing Different Types of Lexical Semantic Information: Evidence From Priming. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **21,** 1-21.

Neely, J.H. (1991). Semantic Priming Effects in Visual Word Recognition: A Selective Review of Current Findings and Theories. In D. Besner & G.W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*, 264-336. Hillsdale, NJ: Erlbaum.

Plaut, D.C. (1995). Semantic and Associative Priming in a Distributed Attractor Network, *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, 37-42. Hillsdale, NJ: Erlbaum.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S. & Patterson, K.E. (1994). Under-standing Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains. *Psychological Review,* to appear.

Plaut, D.C. & Shallice, T. (1993). Deep Dyslexia: A Case Study of Connectionist Neuropsychology. *Cognitive Neuropsychology*, **10**, 377-500.

Schütze, H. (1993). Word Space. In S. J. Hanson, J. D. Cowan & C. L. Giles (Eds), *Advances in Neural Information Processing Systems 5*, 895-902. San Mateo, CA: Morgan Kauffmann.

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

Sejnowski, T.J. & Rosenberg, C.R. (1987). Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, **1**, 145-168.

Shelton, J.R. & Martin, R.C. (1992). How Semantic is Automatic Semantic Priming? *Journal of Experimental Psychology: Learning, Memory and Cognition*, **18**, 191-210.