Connectionist Modelling of Spelling

John A. Bullinaria
Department of Psychology, University of Edinburgh
7 George Square, Edinburgh EH8 9JZ, U.K.
johnbull@uk.ac.ed

Abstract

We present a new connectionist model of human spelling and investigate some of its properties. Although based on Sejnowski & Rosenberg's (1987) NETtalk model of reading, it requires no pre-processing of the training data to align the phonemes and letters. The model achieves 100% performance on the training data (2837 monosyllabic words including many irregular words) and has a generalization performance of about 89%. Under appropriate conditions it exhibits symptoms similar to developmental surface dyslexia and acquired surface dysgraphia. However, its inability to account for phonological dysgraphia and lexical decision leads us to believe that it is a promising candidate for the rule based part of a dual route model but *not* a complete model of spelling on its own.

Introduction

We know from our ability to spell non-words, and also the kinds of spelling errors that humans make on real words, that we must have access to some kind of rule based phoneme to grapheme system (e.g. Kreiner & Gough, 1990). Our ability to spell exception (i.e. irregular) words and to deal with homophones (e.g. 'fare' and 'fair') and to perform lexical decision suggests that there must also exist some kind of lexical/semantic system. How these two processes fit together is not at all clear. The explicit modelling of spelling abilities in humans is thus of importance in understanding the underlying mechanisms involved, e.g. do we require a dual route system with separate semantic and rule-based routes or is a single route system sufficient. It should also provide insight into the spelling difficulties of developmental dyslexics and acquired dysgraphics.

The most realistic connectionist model of spelling to date (Brown, Loosemore & Watson, 1993) was based on the Wickelfeature approach to reading of Seidenberg & McClelland (1989). However, given the intrinsic limitations of Wickelfeature representations, the unacceptably poor generalization ability of this approach and the difficulties involved in interpreting the outputs (Besner et al., 1990), we will adopt a different strategy.

In Bullinaria (1993b, 1994) it was shown how the NETtalk model of Sejnowski & Rosenberg (1987) could be modified to produce a connectionist model of reading aloud (i.e. text to phoneme conversion) that required no pre-processing of the training data. It achieved 100% performance on its training data (2998 monosyllabic words including many irregular words), 98.8% generalization performance (on a standard set of 166 non-words), suggested several possible accounts of developmental surface dyslexia and on damage showed symptoms similar to acquired surface dyslexia. It also correlated well with various naming latency experiments. It failed, however, on the lexical decision task, did not exhibit the pseudo-homophone effect (McCann & Besner, 1987), nor provided any explanation of phonological dyslexia (e.g. Shallice, 1988). It was therefore concluded that the model provided a promising basis for the phonological route of a dual route model of reading but could not be considered to be a complete model of reading on its own.

The task of spelling (i.e. phoneme to text conversion) is clearly closely related to the task of reading aloud and so one might think that given a connectionist model of reading it should be straightforward to construct the analogous connectionist model of spelling. In practice, there is more than a simple inverse mapping involved (e.g. Frith, 1980; Kreiner & Gough, 1990): the rule structures are somewhat more ambiguous, with a large proportion of homophones, and it is harder to organise the alignment of the phonemes and letters in the training data without the need for pre-processing by hand.

The Model

The basic reading model presented in Bullinaria (1993b) consists of a standard fully connected feedforward network with sigmoidal activation functions and one hidden layer set up in a similar manner to the NETtalk model of Sejnowski & Rosenberg (1987). The input layer consists of a window of *nchar* sets of units, each set consisting of one unit for each letter occurring in the training data (i.e. 26 for English). The output layer consists of one unit for each phoneme occurring in the training data (i.e. about 38 units). The input words slide through the input window, starting with the first letter of the word at the central position of the window and ending with the final letter of the word at the central position. Each letter activates a single input unit. If there were a one-to-one correspondence between the letters and the phonemes, the activated output phoneme would then correspond to the letter occurring in the centre of the window. Since there can be a many-to-one correspondence between the letters and phonemes, some of the outputs must be blanks (i.e. no phoneme output). The *alignment problem*, i.e. the problem of not knowing where to insert these blanks

in order to align the letters and phonemes appropriately, was 'solved' in the original NETtalk by hand prior to training. For example, the word 'game' has four possible output targets (i.e. alignments), namely /gAm–/, /gA–m/, /g–Am/ and /–gAm/. But only the one that corresponds to a sensible set of letter to phoneme rules, namely /gAm–/, should be used for the training target. (We use the phoneme notation and conventions of Seidenberg & McClelland, 1989, throughout.)

In Bullinaria (1993b) it was shown that, using a multi-target approach to learning from ambiguous training data (Bullinaria, 1993a), it was possible for the network to *learn* for itself which was the most appropriate alignment and that this alignment was not always necessarily the obvious one that we might choose by hand. The procedure essentially works by considering all possible alignments and for each word allowing the network to train only on the target that already has the lowest error score. This corresponds to the human tendency to look at new learning instances from the point of view that best fits in with our existing knowledge. The word 'game' would be presented four times, once with each of the four letters in the centre of the window. For each presentation there are different possible target phonemes corresponding to the different alignments (e.g. the four possible targets for the 'm' are /m/, /–/, /A/ and /A/ respectively). Summing the output activation error scores over the presentations gives the total error scores for each alignment. The alignment with the lowest total error is then used to train the network in the usual manner. Assuming we keep the learning rate sufficiently low, the fact that the regular letter to phoneme correspondences will naturally tend to dominate the weight changes allows the system to settle into an optimal set of alignments even if we start from random initial weights.

It became clear quite early on that, because certain letters (e.g. 'x' and 'u') sometimes corresponded to more than one phoneme (e.g. /ks/ in 'box' and /yU/ in 'cube'), the standard NETtalk single phoneme output was not allowing optimal alignment. Things are even worse for spelling because (in English) one phoneme (e.g. /O/ and /A/) can correspond to up to four letters (e.g. 'ough' in 'though' and 'eigh' in 'eight'). The obvious solution is to allow more than one phoneme or letter to be output per word presentation, though the number of targets per input can then grow rather large. For example, with just two output phonemes the number of targets for the word 'cube' rises from one (i.e. /kyUb/) to 105 (e.g. /k– yU –b —/). If we allow four output characters per presentation, the number of targets generated becomes prohibitive. One way to restrict the number of targets without making any assumptions about the nature of the training data is for each presentation (corresponding to one input character) to have the set of output characters left justified, i.e. to allow blank outputs only to the right of any phonemes/letters. For the word 'cube' we are thus left with only nineteen targets (e.g. /k– yU b– ––/ is allowable whereas /k– yU –b —/ is not). Proceeding similarly, even the four output spelling model becomes feasible. We thus end up with the network architecture shown in Figure 1. The window size *nchar* is determined by the long range dependencies in the training data and 13 characters was found to be sufficient for our purposes.

There are 427 homophones in our training data, including numerous homophone triples, and it is well known that such ambiguities can cause serious problems with neural network learning and generalization (e.g. Bullinaria, 1993b). In humans we make use of context information to resolve these ambiguities. For the reading model it was shown that a single extra (context) marker appended to one of each of the thirteen pairs of homographs was sufficient. In humans, we obviously use much richer context information than a single marker, yet more markers mean a larger network. For our initial study, we compromised by introducing seven context markers (i.e. characters) and assigned them on a semi-regular basis to the training data to resolve the homophone ambiguities. The final network thus had 13 sets of 38 + 7 input units (for the phonemes plus context markers) and 4 sets of 26 + 1 output units (for the letters plus blank) organised in the same way as the reading model.

Simulation Results

The networks were trained using the back-propagation gradient descent learning algorithm (Rumelhart, Hinton & Williams, 1986) with a training corpus of 2837 monosyllabic words consisting of the original Seidenberg & McClelland (1989) set plus 101 other words missing from that set minus 161 words that had more than 184 targets. (These restrictions were due to the limited memory and

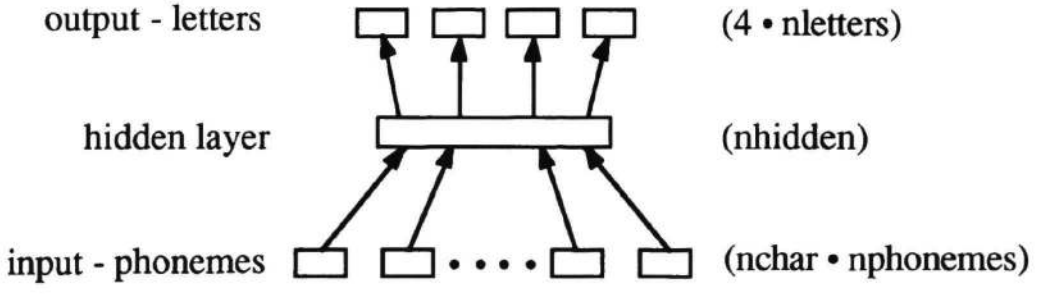

Figure 1: The network architecture for the spelling model.

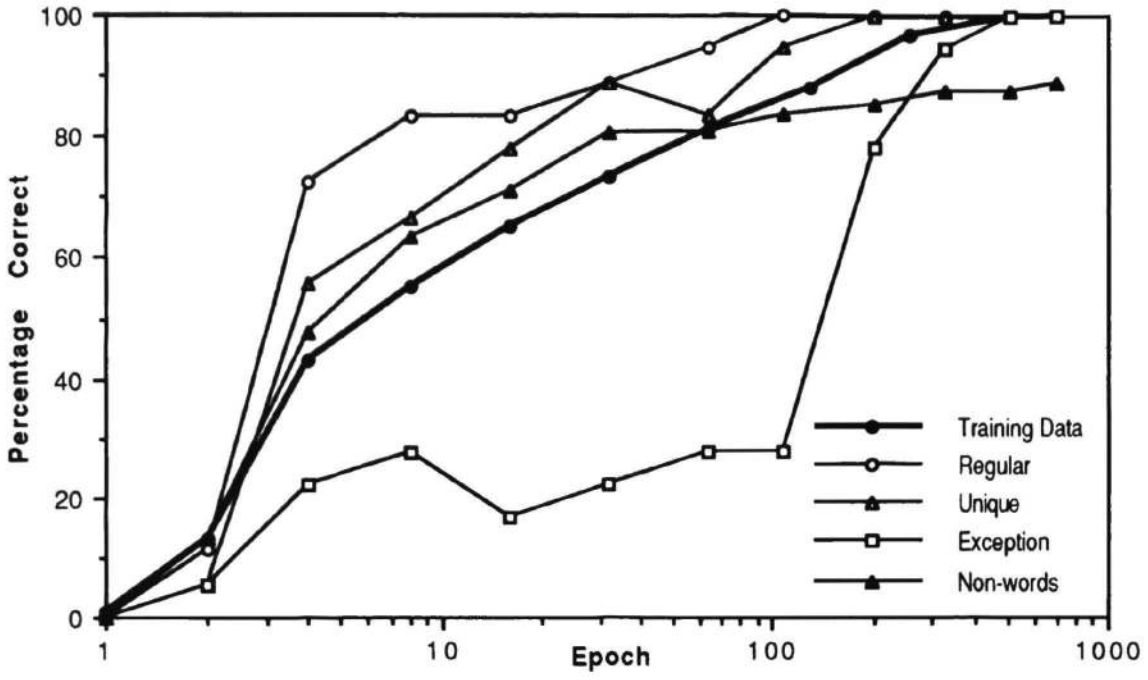

Figure 2: Typical learning curves for the standard spelling model with 300 hidden units.

speed of the computer and the way the algorithm was coded: smaller runs demonstrated that larger numbers of targets, multi-syllabic words and longer range dependencies did not prevent the algorithm from working.) In addition to learning the correct spellings we also expect the model to exhibit various word frequency effects found in humans. Presenting the training data using real word frequencies was not feasible due to the long training times that would be required to learn all the low frequency exception words. About 25% of the training data was therefore presented in each epoch in random order with the same logarithmically compressed frequency distribution as used by Seidenberg & McClelland (1989).

Figure 2 shows the learning curve for one run with 300 hidden units, learning rate of 0.05, momentum 0.9 and sigmoid prime offset 0.1 (Fahlman, 1988). Training was stopped after 700 epochs, by which time the network was achieving 100% performance on the training data. Also plotted on this graph are separate curves for regular, unique and exception words. For ease of comparison with other models we used as closely as possible the matched word sets of Brown et al. (1993). Four of their words were removed because they were absent from the training data ('shrinks', 'breadth', 'owes', 'dozed'), one was removed because it was mis-classified in the dialect of the training data ('huge') and two exception words ('monk', 'sweat') were added to give 18 words in each group. (In terms of rimes we have: Regular = Many friends, no enemies; Unique = No friends, no enemies; Exception = No friends, many enemies.)

The generalization performance was tested on the same set of 166 non-words traditionally used to test reading models: the regular non-words and exception non-words of Glushko (1979, Experiment 1) and the control non-words of McCann & Besner (1987, Experiment 1). The pronunciations were derived from those of the base words in the original experiments. Some of the Glushko non-words (e.g. /fEl/ → 'feal') were homophonic with words in the training data (i.e. 'feel') and these had their first letter changed (i.e. giving /gEl/ → 'geal') so that no non-word pronunciations had been used for training. The acceptable spellings were derived by matching common word segments in the training data. This generally led to several allowable spellings for each input pronunciation, e.g. /hEf/ → 'heef' as in 'beef' and 'heaf' as in 'leaf'

The final generalization performance of 88.6% is quite poor compared to the corresponding reading model (which achieved up to 98.8%). However, considering the *ad hoc* procedure used to deal with the homophones and the limited amount of training data, it is encouragingly high. It is also somewhat more than we could expect to obtain with Wickelfeature based models which only achieve about 60% on the simpler reading case (Besner et al., 1990). The phoneme-letter alignments were checked and found to be very good (e.g. /drWt/ → 'd--- r--- ough t---') and consistent, but not always as one might expect (e.g. /dok/ → 'd--- oc--- k---'). As with many models of this sort, the errors can reveal more about the underlying mechanisms than the correct responses. Of the 19 errors (shown in Table 1), none were totally wrong and 11 had close output rivals which were acceptable spellings. Others, such as 'praine', 'doade' and 'vox', may be considered acceptable under more generous scoring criteria. It is also interesting to note that 8 of the erroneous outputs corresponded to actual words in the training data. Given the nature of these generalization errors, we see that the network actually does somewhat better than the 88.6% correct figure suggests.

It has been suggested (Seidenberg & McClelland, 1989) that the network output activation error scores should be

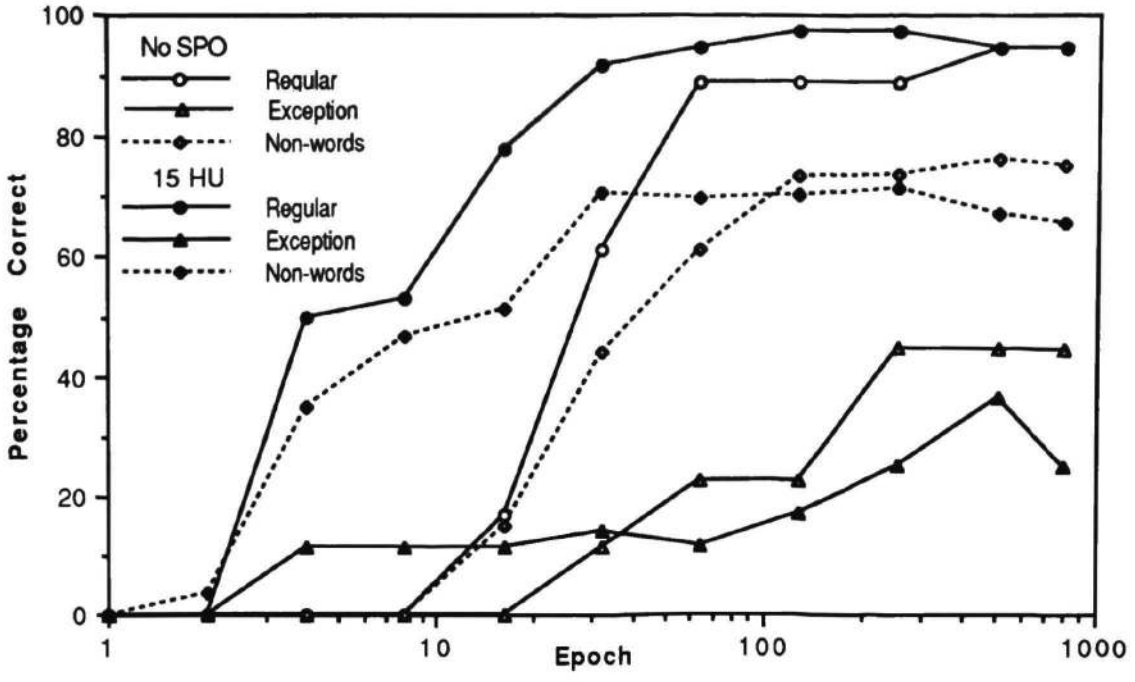

Figure 3: Two sets of learning curves reminiscent of developmental dyslexia.

monotonically related to human response times and this has indeed been found to be the case for the reading model (see Bullinaria, 1994, for a more detailed discussion). For the error scores of our fully trained spelling model we find a significant type effect between the regular and exception words ($F(1,32) = 21.6$, $p < 0.0001$) and between the unique and exception words ($F(1,32) = 29.1$, $p < 0.0001$) but no significant effect between the regular and unique words ($F(1,32) = 0.1$, $p = 0.74$) nor any frequency nor interaction effects ($p > 0.1$ in all cases). The lack of frequency effects is almost certainly due to our logarithmic compression of the word frequencies (which was necessary for the simulations to finish in a reasonable time). Results from the corresponding reading model (Bullinaria, 1994) indicate that frequency effects should, in principle, arise in this kind of model. The problem of modelling frequency effects is clearly something that must be addressed more carefully in future work in this field.

INPUT	OUTPUT	CLOSE RIVALS
prAn	praine	
woS	wash	
s^f	sugf	suff, sugh
grUl	growl	gruwl
pOm	pom	pome
sud	soold	sood
h^v	have	
bIld	build	bild, biled
wuS	wosh	woosh
tul	toll	tool
dut	dot	dut, doot
nent	kent	knent, nent
pId	pid	pide, pied
gof	golf	
lOks	lox	
kEr	ceer	
v*ks	vox	
dOd	doade	
tUd	tuod	teod, tood

Table 1. Non-word spelling errors

Brown et al. (1993) found a significant difference at all stages of learning between the regular and unique words both in terms of error scores and number of errors. However, Brown et al.'s model splits each word into triples of characters and so the regularity and frequency of word endings are important in the learning process. Our model works with only one phoneme at a time and so learning only tends to be slowed by lack of word ending friends when one of the phonemes (usually a vowel) is irregular. The unique words used here are in fact very regular and hence the lack of difference. It is interesting to note, however, that we do get significant type differences when the network resources are limited. For a network with only 40 hidden units (which is only just enough to get perfect performance on the training data) we found significant ($p < 0.003$) differences for all three type pairs. This might explain any differences found in humans (Brown et al., 1993), but they could just as easily be due to something other than the use of simple phoneme to grapheme rules: for example, humans may acquire a lexicon of common sub-words that can be spelt directly, common sub-words may be recognised more quickly, etc.

Finally, we note that although there is a significant increase in output error scores for non-words compared with words in the training data, there is considerable overlap in

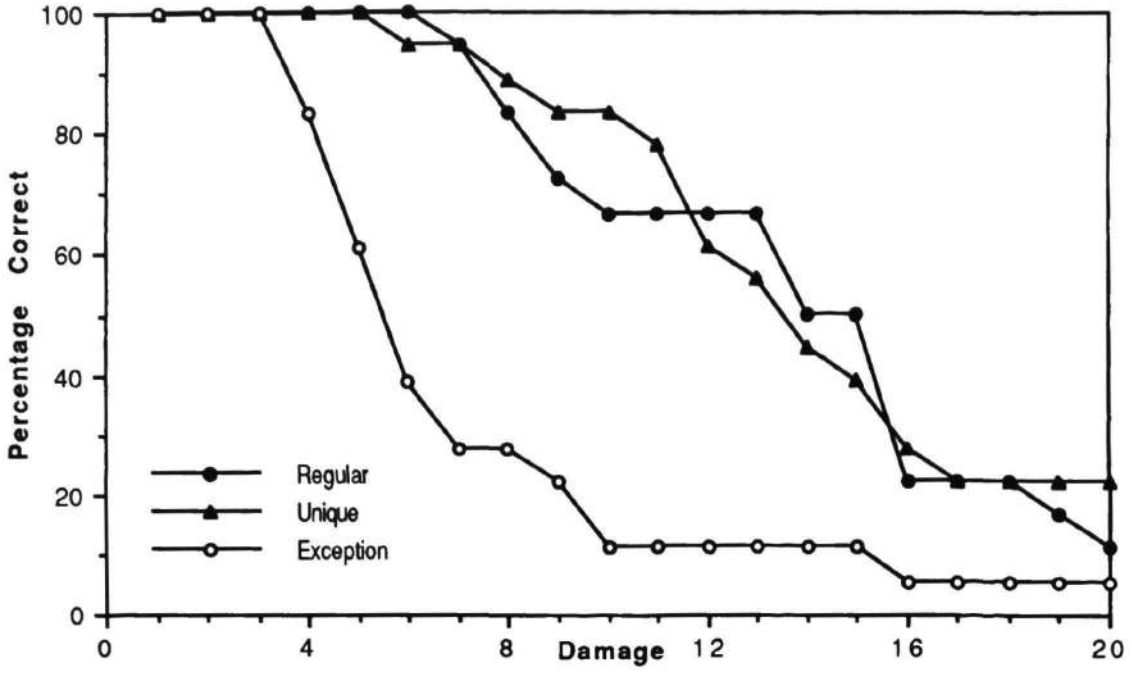

Figure 4: Performance after damage by weight scaling of the standard spelling model.

scores between the two groups, so this *cannot* be used as a basis for lexical decision.

Developmental Dyslexia

Any realistic model of spelling should be able to account for the dissociation between the learning of regular and irregular words found in many developmentally dyslexic children (e.g. Frith, 1985). It is clear that if the computational resources are lowered enough (e.g. by reducing the number of hidden units or connections and hence the effective number of free parameters) then the system will fail to learn properly. Figure 3 shows that using only 15 hidden units we do get a clear dissociation between the learning of the different word types (averaged over two runs). We can expect similar patterns of performance with other forms of resource restrictions. A second possible cause of dissociation is related to our use of a sigmoid prime offset (SPO) in the learning algorithm to prevent the output activations getting stuck hard wrong (where the error propagated back is zero). The learning curves for the 40 hidden unit case with no SPO are also shown in Figure 3. With an SPO we achieve 100% performance for all word types by epoch 700. We are not suggesting that real brains employ an SPO, but simply that spelling difficulties could equally well occur from particular learning problems within our model as from general limitations on the computational resources.

Note that the final proportions correct for the regular and exception words are similar to those found at earlier stages in normal development. This is predictable from the nature of the learning algorithm: the rules are learnt first and (if sufficient resources remain and the learning algorithm is still capable of learning) the exceptions later. Our model thus provides equally good "deviant" and "delayed" accounts of developmental dyslexia (e.g. Frith, 1985). There is some evidence of more than proportional reduction of performance on non-words in dyslexics (e.g. Frith, 1980) and unique words in our reduced resources models, but we consider it premature to say anything definite about this here (cf. Brown et al., 1993).

Acquired Dysgraphia

An important constraint on cognitive models is provided by the performance of the model after various forms of damage (e.g. Shallice, 1988). In fact, one of the main reasons for believing in the dual route model of spelling (and reading) is provided by the performance of patients after certain forms of brain damage: Surface dysgraphics have difficulty with spelling irregular words compared with regular words while phonological dysgraphics can spell both regular and irregular words but lose the ability to spell non-words.

These two forms of dysgraphia are usually explained by a dual (rule plus lexical/semantic) route model by losing one of the two routes but not the other. Given that our model has no way to learn a lexicon or semantics and is unable to perform lexical decision, we cannot really expect it to exhibit phonological dysgraphia. However, even if some form of separate lexical/semantic route exists in addition to the rule-based route modelled here, our model must still be able to exhibit surface dysgraphia when damaged: Since our model can spell both regular and irregular words, it cannot simply be a matter of losing the lexical/semantic route alone.

A range of forms of damage that may be inflicted on models such as ours were discussed in Bullinaria (1994). The five main types are the global reduction of all weights

by constant amounts, the global scaling of all weights by constant factors, the addition of Gaussian random noise to all weights, the random removal of connections and the random removal of hidden units. For networks with large numbers of hidden units, globally reducing all the weights by successive factors of 0.95 is a convenient deterministic procedure that has a very similar effect to adding random noise or removing appropriate fractions of the hidden units and/or connections at random. Figure 4 shows that the effect of doing this does give the required dissociation. Moreover, many of the errors are regularisations (e.g. /sed/ → 'sed' not 'said' and /giv/ → 'giv' not 'give') as is found in human surface dysgraphics. Similar patterns of errors are obtained by explicit application of the other four forms of damage.

If our model constituted just one route of a dual route model, it is possible that the lexical/semantic route alone could learn to deal with all the exception words before our rule-based route had time to finish mastering them. Overall efficiency might then mean that learning in the rule based route need not proceed beyond a certain stage (e.g. somewhere between epoch 30 when the generalization performance begins to level off and epoch 120 when the exception performance becomes significant). If this happens, our rule based route already has a ready made dissociation for when the lexical route is lost and we may explain surface dyslexia in that way. This, of course, is how things are usually explained in the *traditional* dual route model.

Discussion

We have presented a simple connectionist model of spelling that accounts for several effects found in humans. Moreover, unlike other models, it can handle multi-syllabic words and long range dependencies yet requires no pre-processing of the training data. Future simulations, with more hidden units, more hidden layers, larger training data sets (including many multi-syllabic words) and stress markers (as in Sejnowski & Rosenberg, 1987), should show even better performance.

However, certain fundamental features of the model remain unsatisfactory. Primarily, we know from its inability to account for phonological dysgraphia and lexical decision that the model is lacking a lexical or semantic route which is also necessary to tie in with semantic processing, etc. Progressing from here to include such a sub-system will also allow us to employ a more principled mechanism to deal with homophones, which in turn should further improve the generalization performance. Another area in need of improvement is in the use of the moving window approach, which is not very biologically nor psychologically realistic in itself and is likely to prove difficult to coordinate with a separate lexical/semantic route. In principle, a series of recurrent connections, such as used by Jordan (1986), could replace the moving window as a method of dealing with the context information. Work is currently in progress to re-implement the model in terms of a network with recurrent connections and a rudimentary lexical/semantic route.

References

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

Brown, G.D.A., Loosemore, R.P.W. & Watson, F.L. (1993). Normal and dyslexic spelling: A connectionist approach. Technical report, University of Wales, Bangor.

Bullinaria, J.A. (1993a). Neural Network Learning from Ambiguous Training Data. Submitted to *Connection Science*.

Bullinaria, J.A. (1993b). Neural Network Models of Reading Without Wickelfeatures. *Proceedings of the 2nd Neural Computation and Psychology Workshop*, Edinburgh.

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

Glushko, R.J. (1979). The Organization and Activation of Orthographic Knowledge in Reading Aloud. *Journal of Experimental Sciences: Human Perception and Performance*, **5**, 674-691.

Fahlman, S.E. (1988). Faster-Learning Variations on Back-Propagation: An Empirical Study. In *Proceedings of the 1988 Connectionist Models Summer School*, Morgan Kauffmann.

Frith, U. (1980). *Cognitive Processes in Spelling*. London: Academic Press.

Frith, U. (1985). Beneath the Surface of Developmental Dyslexia. In *Surface Dyslexia* (eds. K.E. Patterson, J.C. Marshall & M. Coltheart). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jordan, M.I. (1986). Attractor Dynamics and Parallelism in a Connectionist Sequential Machine, *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 531-536). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kreiner, D.S. & Gough, P.B. (1990). Two Ideas about Spelling: Rules and Word-Specific Memory. *Journal of Memory and Language*, **29**, 103-118.

McCann, R.S. & Besner, D. (1987). Reading Pseudohomophones: Implications for Models of Pronunciation Assembly and the Locus of Word-Frequency Effects in Naming. *Journal of Experimental Psychology: Human Perception and Performance*, **13**, 14-24.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning Internal Representations by Error Propagation, in *Parallel Distributed Processing*, Volume 2 (eds. D.E. Rumelhart & J.L. McClelland). Cambridge, Mass: MIT Press.

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.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.