APHASIOLOGY, 2005, 19 (9), 789–806

Psychology Press
Taylor & Francis Group

Using computational, parallel distributed processing networks to model rehabilitation in patients with acquired dyslexia: An initial investigation

Stephen R. Welbourne and Matthew A. Lambon Ralph
University of Manchester, UK

Background: Traditional cognitive neuropsychological models are good at diagnosing deficits but are limited when it comes to studying recovery and rehabilitation. Parallel distributed processing (PDP) models have more potential in this regard as they are dynamic and can actually learn. However, to date very little work has been done in using PDP models to study recovery and rehabilitation.
Aims: This study seeks to demonstrate how a PDP model of acquired dyslexia can be extended to provide a computational framework that is capable of making predictions about the relative effectiveness of therapeutic interventions.
Methods & Procedures: A replication of Plaut, McClelland, Seidenberg, and Patterson's (1996, simulation 2) model of word reading was trained and then damaged. This damaged network was then retrained in a number of different ways designed to model both natural (spontaneous) recovery and recovery that can be attributed to a specific therapeutic intervention.
Outcomes & Results: Interventions that used regular words were more effective than interventions based on inconsistent words. Early intervention (during the period of spontaneous recovery) was more effective than late intervention.
Conclusions: These results suggest that this technique has the potential to provide a useful input to the therapeutic arena. The potential opportunities for further work are discussed.

Despite considerable early promise, cognitive neuropsychology has so far been somewhat disappointing in its ability to contribute to our understanding of rehabilitation. At first sight this seems rather surprising as cognitive neuropsychology is essentially concerned with using data from brain-injured patients to construct models of normal cognitive function. It should, therefore, be well placed to address issues of therapy design. In the early years of the discipline there was considerable optimism that cognitive neuropsychology would have a major part to play in the field of rehabilitation (see, for example, Caramazza & Hillis, 1993; Seron & Deloche, 1989) and indeed there were a few early success stories (Coltheart & Byng, 1989; De Partz, 1986). However, the early optimism has largely given way to pessimism; when a recent special edition of *Neuropsychological Rehabilitation* was devoted to the role of cognitive neuropsychology in language rehabilitation, the overall tone of the papers varied between muted pessimism (Shallice, 2000) and outright scepticism (Basso & Marangolo, 2000).

Address correspondence to: Stephen R. Welbourne, School of Psychological Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK. Email: Stephen.R.Welbourne@man.ac.uk

This research was supported by grants from NIMH (No. P50-MH64445), BBSRC (S20390) and the University of Manchester.

http://www.tandf.co.uk/journals/pp/02687038.html DOI:10.1080/02687030500268811

Traditionally, cognitive neuropsychology has operated by constructing models of cognitive processes from information about the patterns of breakdown in brain-damaged patients. These models are static in nature, consisting of a set of boxes connected by arrows where the boxes tend to represent stored representations and the arrows represent the processes that map from one representation to another. These models have been successful in categorising neuropsychological deficits and, in a number of areas, have challenged explanations in terms of broad syndromes. One good example of this is the breakdown of the broad syndrome of acquired dyslexia into three central dyslexias (surface, phonological, and deep) and four peripheral dyslexias (neglect dyslexia, visual dyslexia, attentional dyslexia, and letter by letter reading). While this approach has been successful in terms of description and diagnosis, its success in understanding rehabilitation has been limited. It is often suggested that the reason for this may be that these types of models do not incorporate any mechanism for learning (Baddeley, 1993).

One possible alternative to traditional box-and-arrow models is connectionist or parallel distributed processing (PDP) modelling. PDP models consist of interconnected sets of neurone-like units. The connections between the units have adjustable weights in the same way that human synaptic weights are variable. For both the models and humans the knowledge of the system is encoded in this pattern of weights. If a significant number of weights or units are damaged then the model's performance is impaired. Like human performance after brain damage, this impairment is gradual and increases in tandem with the severity of the damage.

PDP models learn to perform mappings between different domains; reading aloud is encapsulated by the mapping between orthographic and phonological patterns. This type of reading model learns through an exposure to an environment that consists of input patterns (orthographical representations of words) and target patterns (phonological representations of words). Over repeated exposure to this environment the weights, in the model, are adjusted in such a way as to bring the output of the model gradually closer to the target. The fact that these networks do actually learn means that they offer much greater potential as models of rehabilitation than more traditional approaches. Indeed this kind of modelling has the potential to encompass a complete cycle of development, acute damage, spontaneous recovery, and rehabilitation.

There has been much debate as to the status of PDP modelling within cognitive neuropsychology. Indeed, the role of connectionism was one of the key themes to emerge in a recent issue of the journal *Cognitive Neuropsychology* (Vol. 21, Issue 1), devoted to the future of the discipline. In the target article, Harley (2004) lamented the fact that PDP techniques are too often ignored. He argued that PDP models are important because they focus on the process of cognition in a realistic way using neural-like parallel processing. Many of the contributing authors were also sympathetic to this viewpoint, although both Coltheart (2004) and McCloskey (2004) argued against it. Rightly, in our view, Dell (2004) noted that this technique has a number of benefits though one should never expect computational models to be perfect in all respects. Lambon Ralph (2004) suggested that one of the strengths of the PDP approach lies in its ability to specify both function and architecture explicitly and that, often, this leads to a more parsimonious cognitive model than the equivalent box-and-arrow approach.

Evidence for the significant potential of PDP modelling can be observed through a number of very influential models capturing data from many different domains of human performance. As well as models of reading, which we shall discuss in more detail later (Harm & Seidenberg, 1999, 2004; Plaut et al., 1996; Seidenberg & McClelland, 1989), there are models of past-tense generation (Joanisse & Seidenberg, 1999; Rumelhart &

McClelland, 1986), memory (McClelland, McNaughton, & O'Reilly, 1995), face recognition (Burton, Bruce, & Johnston, 1990), speech production (G. S. Dell, Schwartz, Martin, Saffran, & Gagnon, 1997), semantics (Rogers et al., 2004), and no doubt many more (the list is indicative rather than exhaustive). In all of these cases the models are actually reproducing some aspect of human behaviour which can be directly compared against that of the relevant human population. This means that these computational models are open to much more rigorous testing than could ever be applied to box-and-arrow models.

Although there are many successful computational models, very few of them address the issues of recovery or rehabilitation; neither did these issues feature in the published debate reported above. This omission is surprising in view of the obvious suitability of PDP modelling for this kind of study (learning and relearning are intrinsic to such models). We suspect that this reflects a combination of the traditional focus of cognitive neuropsychology on disorders rather than on recovery, and the sheer computational difficulty of running models of rehabilitation.

In this paper we describe a set of PDP simulations of a reading model that builds on previous work by ourselves and others. In order to understand the context of this work with respect to previous models of reading we present a brief history below.

Probably the first PDP reading model of any serious merit was Seidenberg and McClelland's (1989) model of word reading, hereafter known as SM89. SM89 is a restricted implementation of a larger triangle model. While the triangle model envisages that information will be simultaneously processed by a phonological and a semantic pathway, SM89 implemented only the phonological portion of this model (see Figure 1). Despite this, SM89 was able to learn the pronunciations of 2820/2897 monosyllabic words (97.3%). These included irregular words like ''pint'' and ''yacht'', which according to the dual route model require processing by an exclusively lexical route. What is more, the model displayed the same kind of frequency/consistency interaction that is found in normal readers—high-frequency words all had quite low error scores but inconsistent low-frequency words had a much higher error rate than did low-frequency consistent words. The model was also able to read some nonwords. However, in this

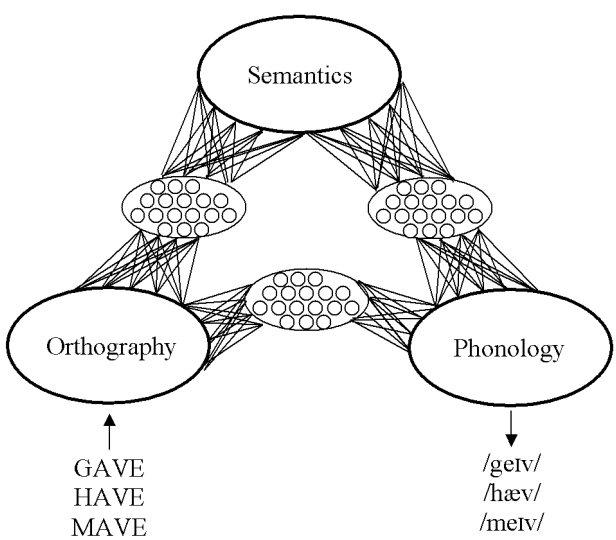

Figure 1. The triangle model from Seidenberg and McClelland (1989).

respect it performed considerably worse than normal human performance; Besner, Twilley, McCann & Seergobin (1990) reported that on a regular nonword list taken from Glushko (1979) the model was only 59% correct whereas normal readers are 94% correct.

Although there was some match between the performance of the model under damage and that found in surface dyslexia, it was not very compelling (Patterson, Seidenberg, & McClelland, 1989; Plaut et al., 1996). The classic symptoms of surface dyslexia are poor reading of low-frequency irregular words and a strong tendency for errors on these words to be regularisations (e.g., reading PINT to rhyme with MINT). In this case, the model did not display a sufficiently large consistency/regularity interaction when damaged (it made errors on all classes of words not just the low-frequency irregular ones), neither did it produce enough regularisation errors to make a convincing case that it was modelling surface dyslexia. In summary, SM89 was inadequate on two major counts; the undamaged model did not read nonwords accurately enough and the damaged model did not display a sufficiently large frequency/consistency interaction.

Plaut et al. (1996) returned to the issue of modelling reading and surface dyslexia. They made a close analysis of the performance of SM89 and concluded that many of its weaknesses stemmed from the kind of representations that it utilised. In a new version of the reading model, they adopted representations that were better able to capture regularities in the mapping between orthography and phonology. This model, hereafter PMSP96, was trained on a set of 2972 monosyllabic words similar to those used in SM89. Again, like SM89, it was able to learn the correct pronunciation of all the words in the training corpus and it exhibited the standard frequency/consistency effects on reading times. Unlike SM89, it also performed at human levels in nonword reading; it could correctly read 97.7% of the consistent nonwords from Glushko (1979). When the model was lesioned, however, it did not perform in a way that mimicked surface dyslexia—although it was considerably closer to it than SM89. In particular, those lesions that produced the correct level of impairment on high-frequency irregular words did not sufficiently impair low-frequency irregular words. Also, the level of regularisation errors, while higher than that found in SM89, did not correspond to that found in dyslexic patients. Plaut et al. considered that these results implied that surface dyslexia could not be modelled within a purely phonological model and that some consideration of the role of semantics was required. They went on to demonstrate that the pattern of deficits found in surface dyslexics could be very effectively modelled by training the network in the presence of a semantic input and then removing that semantic contribution—an association that is found in patients with semantic dementia (Patterson & Hodges, 1992).

Despite the success of PDP models of reading there has been very little work to date on recovery following damage. The few studies that have been conducted have concentrated solely on the translation of orthography to phonology via semantic representations (the semantic route). Early work by Hinton and Sejnowski (1986) demonstrated that retraining was faster than the original learning and provided some evidence that retraining on a subset of items could generalise to the rest. More recently, Plaut (1996) used a version of his deep dyslexia model (Plaut & Shallice, 1993) to investigate retraining. Again, he found retraining to be faster than original learning. He also demonstrated that recovered performance was dependent on the location of the damage and on the typicality of the items used in retraining—atypical items provided better generalisation than did typical ones. These models differ from the current simulations in two critical respects. First, previous models adopt the premise that the damaged state of the model should be analogous to the performance of brain-damaged patients after the period of spontaneous recovery. Thus, previous work has focused on modelling

rehabilitation and, in particular, the ability of models to generalise from retraining on a reduced training corpus to the remaining items. By contrast, Welbourne and Lambon Ralph (2005) suggested that the lesioned model might be analogous to a patient immediately after brain damage and that retraining after damage (using the full training corpus) would be the equivalent to spontaneous recovery. Their paper sought to illustrate the importance of neural plasticity to understanding patterns of impaired cognitive performance.

The theoretical position behind that paper is held in common with this study and revolves around the proposition that recovery after brain damage may be, at least in part, attributable to synaptic weight changes rather than purely physiological factors: If the human brain's ability to perform accurately depends on the pattern of synaptic weights then it seems reasonable to assume that the removal of a proportion of those weights will not leave the remaining synapses optimally configured to perform the task. Further, provided that there exists some optimisation process by which the synaptic weights can change (learning) then it seems inevitable that some of the recovery that we observe in patients after brain damage must be attributable to synaptic change. This kind of mature synaptic plasticity has been studied mostly in the context of cortical sensory maps (for a review see Buonomano & Merzenich, 1998) and it is clear that these maps are capable of undergoing extensive modification, presumably as a result of some learning process operating at the synaptic level.

Welbourne and Lambon Ralph (2005) explored this hypothesis by demonstrating that, when a model of reading was damaged and allowed to relearn, its performance improved considerably. Furthermore, it emerged that whilst the performance immediately post-damage was relatively undifferentiated (regular and irregular words read with similar accuracy) once the model had been allowed to relearn, its behaviour began to resemble that of surface dyslexic patients. In the light of this finding, Welbourne and Lambon Ralph suggested that this post-damage learning period might be equivalent to the period of spontaneous recovery seen in patients, and that synaptic changes occurring during this period might be critical in understanding the pattern of chronic performance in patients.

In this paper we take the model one stage further by using it to investigate the effect of specialised rehabilitation therapy in conjunction with a period of spontaneous recovery. We seek to answer some basic yet fundamental questions concerning the effectiveness of therapeutic intervention:

(1) Does therapeutic intervention using a small number of words have the potential to generalise to untrained items?
(2) Does the type of training stimuli affect subsequent performance?
(3) Is therapeutic intervention still effective over and above the effect of spontaneous recovery?
(4) Is early or late intervention more effective?

To answer questions 3 and 4 (simulation 2) we need some kind of operational definition of how rehabilitation therapy and spontaneous recovery differ in terms of what is happening to the network. We adopted a simple definition that should be applicable across a range of domains. For spontaneous recovery we assumed that there will be some re-exposure to the original learning environment. However, we anticipated that the salience of this re-exposure would be low—in our society it is hard to avoid words; they appear all around, often in contexts that give strong clues to meaning, but it is also easy to ignore them. We modelled this by re-exposing the network to words randomly selected from the

original corpus but only allowing a very small amount of learning to occur for each exposure. By contrast rehabilitation was defined as training on a small set of words but with a very high salience—patients in rehabilitation are usually very motivated to succeed. This was modelled by selecting sets consisting of 34 therapy items and training the network on each item in turn, with much more learning allowed during this type of exposure to the learning environment.

METHODOLOGICAL DETAILS

Overview

Starting from a replication of Plaut et al.'s (1996) simulation 2 (also described in Welbourne & Lambon Ralph, 2005) two main simulations were conducted. The purpose of the first simulation was to investigate the overall potential of rehabilitation therapies, based on a small subset of the original corpus, to generalise to the rest of the corpus. The second simulation was designed to look in more detail at the interaction between therapeutic intervention and spontaneous recovery. It also aimed to establish whether there was any difference between early or late intervention. The same training stimuli and measure of network performance were used in both simulations. Accuracy was measured in terms of the percentage of words from the original corpus that were read successfully excluding the items within the therapy sets. In all simulations the results were averaged across 10 different random lesions of 10 networks trained with different random initial weights. Thus each data point represents the average of 100 simulation runs.

Network architecture and learning algorithm

The architecture, training, and representations used in this simulation were chosen to be as similar as possible to that used by Plaut et al. (simulation 2, 1996).[1] Each of these key features is summarised below. Figure 2 shows the architecture of the network that was used throughout this set of simulations. There were three sets of units: 105 grapheme units, 100 hidden units, and 61 phoneme units. Each layer of units was fully connected to the next layer up. Thus every grapheme unit was connected to every hidden unit and every hidden unit was connected to every phoneme unit. The activity level of each unit was set to vary between 0 and 1 as a nonlinear (logistic) function of the unit's total input.

The initial weights on the connections were set to random values between -0.1 and $+0.1$. The network was then trained using the standard backpropagation learning algorithm with momentum enabled only if the gradient of the error slope was less than 1. Like PMSP96, cross entropy was used as the error function. The learning rate for the network was set to 0.05 and the momentum was 0.9.

It should be noted that this learning procedure differs slightly from that used in PMSP96 where each connection was allowed to modify its own learning rate in a procedure known as delta bar delta learning. The procedure used here, however, is computationally simpler and results in very slightly better performance than was found in PMSP96.

Orthographic and phonological representations

The network used the same representations as PMSP96. These representations divide each word into three parts (onset, vowel, and coda) and then use specific units to code for particular graphemes or phonemes occurring within each part. In addition, the phono-

[1] We are grateful to David Plaut for sharing his training patterns with us.

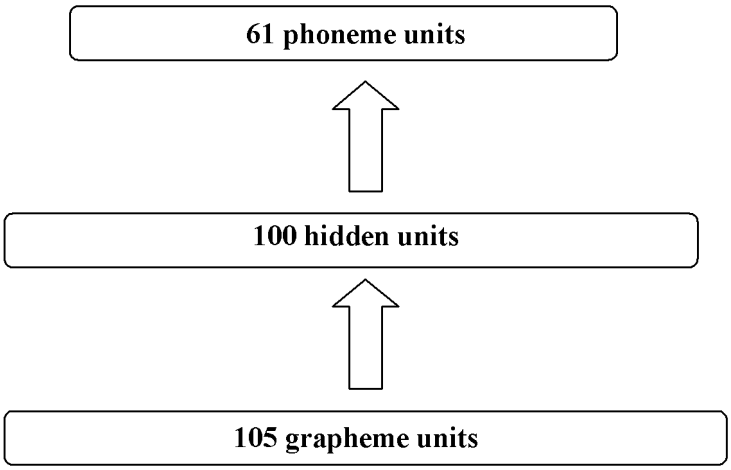

Figure 2. Network architecture.

logical onset and coda are further divided into groups of mutually exclusive phonemes so that, when reading off the unit activations, only the most active member of each group is a candidate for inclusion in the output phoneme string. Table 1 shows the representation scheme used in this simulation (phonological subgroups are separated by extra spaces). In general, words are coded from left to right so that if more than one unit is active in the onset or coda then the output is read in the order that they appear in the table. The only exception to this occurs for the phonemes pairs p–s, k–s, and t–s, which can occur either way round in the phonological coda. To cater for this, special units ks, ps, and ts are used to determine the order. If both s and p are active, then they are taken in the order sp unless the ps unit is active, in which case the order is reversed.

INITIAL TRAINING PROCEDURE

The network was trained using full batches with the same corpus of 2998 monosyllabic words used in PMSP96. The frequency (Kuçera & Francis, 1967) of each word was used to scale the error derivatives for the purposes of backpropagation. This has the same effect as using real frequencies to determine the probability of a word being presented for

TABLE 1
Orthographic and phonological representations

	Orthographic units
Onset	Y S P T K Q C B D G F V J Z L M N R W H CH GH GN PH PS RH SH TH TS WH
Vowel	E I O U A Y AI AU AW AY EA EE EI EU EW EY IE OA OE OI OO OU OW OY UE UI
Coda	H R L M N B D G C X F V J S Z P T K Q BB CH CK DD DG FF GG GH GN KS LL NG NN PH PP PS RR SH SI SS TCH TH TS TT ZZ U E ES ED
	Phonological units
Onset	s S C z Z j f v T D p b t d k g m n h l r w y
Vowel	a e i o u @ ^ A E I O U W Y

training; however, it has the considerable advantage that every word can still be presented once every epoch, thus considerably compressing the required training time (see Plaut et al., 1996, for a fuller discussion of this issue). To eliminate the possibility that the results might be a consequence of one particular set of initial weights, the network was trained 10 times; each time using a different random set of weights as the starting point. These 10 trained networks then formed the starting point for further investigations.

Testing procedure

The performance of the network on all of the words in the original training corpus was tested periodically throughout the training and retraining period. The procedure for determining the phonological output of the network was slightly complicated by the implicit assumptions in the representations used. For the onset and coda, the phonological output was taken to be composed of the most active phoneme in each phoneme group, provided that its activation was greater than 0.5. However, for the vowels, the most active vowel unit was taken as the output even if its activation was less than 0.5. The order of the phonemes was taken from the order in which they appear in the representation. For the pairs of phonemes ks, ps, and ts, this order was reversed (in the coda) if the special ks, ps, or ts units were activated.

In addition to the performance on the training corpus, the percentage of regularisation errors made by the network was also recorded. This was calculated by matching the actual output of the network against a list of possible regularisations for a subset of irregular words drawn from the training corpus. For most words this list consisted of just one possible regularisation but some words (e.g., FLOOD or LOSE) can be regularised in two different ways. The subset of irregular words as well as the list of pronunciations treated as regularisations was taken from those used in PMSP96.

Rehabilitation training sets

Four sets of rehabilitation stimuli were selected from the overall training corpus, with a fifth set consisting of 34 nonwords (not in the original corpus) selected from Glushko (1979). The four sets of words were selected to be split by regularity and frequency. A full listing of all the words and nonwords used in the rehabilitation sets can be found in the Appendix.

Test stimuli

For the majority of testing, the stimuli consisted of the full original corpus minus those words used in the rehabilitation training sets. On a few occasions it was found useful to breakdown the results by frequency and regularity; in these cases stimuli sets consisting of 24 words were used. These were identical to the sets of words used to test PMSP96. None of the words in these testing sets overlapped with any of the words from the training sets. In addition, the training set of nonwords was also used sometimes as a testing set.

Replication of PMSP96

By epoch 1000 the network was performing optimally for all of the stimuli sets. At this point it correctly pronounced all of the words in the corpus with the exception of the homographs and the word ‘‘gent’’, which was mispronounced with a hard g (as in ‘‘gecko’’) on just one of the ten trials. The errors occurring on the homographs are to be expected since they will always occur in single-word reading if no context is provided.

Ignoring the errors stemming from the homographs, the model was performing at 99.97% accuracy, very slightly better than the performance achieved by PMSP96, which was 99.8% correct when trained with real word frequencies. For nonword reading the model was correctly reading 97.0% of the regular nonwords. This is not quite as good as the 97.7% achieved by PMSP96, but it is well within the range of normal human performance, which averages 93.8% (Glushko, 1979). In addition, the error scores showed the expected frequency consistency interaction.

SIMULATION 1: INVESTIGATING THE POTENTIAL OF REHABILITATION TO GENERALISE TO UNTRAINED WORDS

This simulation was designed to investigate the capacity of a damaged network to recover when retrained on only a small subset of the original corpus. A replication of PMSP96 was damaged by removing 60% of the hidden units. The network was then retrained using one of the five rehabilitation-training sets. Training consisted of full batches not weighted by frequency. The learning rate was 0.05 and momentum was 0.9.

Results

Figure 3 shows the results of this rehabilitation training in terms of its effect on the network's overall reading accuracy for the untrained set of words. It is clear that rehabilitation training has considerable potential to generalise to the untrained set of words. For all training sets there is a considerable improvement in reading accuracy over the first 60 epochs. Over this period, reading accuracy moves from about 5% to between 35% and 55 % correct. The high- and low-frequency regular words give the best overall performance with almost identical accuracy rates of approximately 55%. The nonwords are the next best with an accuracy rate of 50%. Both of the irregular training sets do markedly worse. The low-frequency irregulars give an overall accuracy rate of 45%, while the high-frequency irregulars give an accuracy rate of 38%. These irregular sets also have a marked performance peak, which was not evident for the regular training sets. After 80 epochs of retraining, the network's accuracy starts to decline steadily with further retraining, whereas for the regular training sets the network's performance remained stable once the best performance level was reached.

The results for the irregular words are something of a puzzle, as it is not clear why retraining on the low-frequency irregular words should be more beneficial than retraining on the high-frequency ones. There is no frequency weighting involved in the retraining and it is hard to imagine a mechanism whereby frequency weightings from the original training would act to favour retraining on low-frequency words.

One possibility is that the low-frequency irregular word training set is actually slightly more consistent than the high-frequency irregular training set, so that the advantage for low-frequency irregulars is merely an artefact arising from different levels of consistency in the two sets. Unfortunately, there is no method of quantifying degrees of consistency, so it is very hard to control for this possibility. However, within the artificial environment of a network model it is possible to determine experimentally whether the supposed frequency effects can be attributed to differing levels of consistency. To do this a new network was trained in exactly the same manner as before with the one exception that there was no frequency weighting on the initial training. Thus in this case all words were trained with equal weight in both the initial training and rehabilitation phases. The performance of the network at epoch 80 was measured and compared to the performance

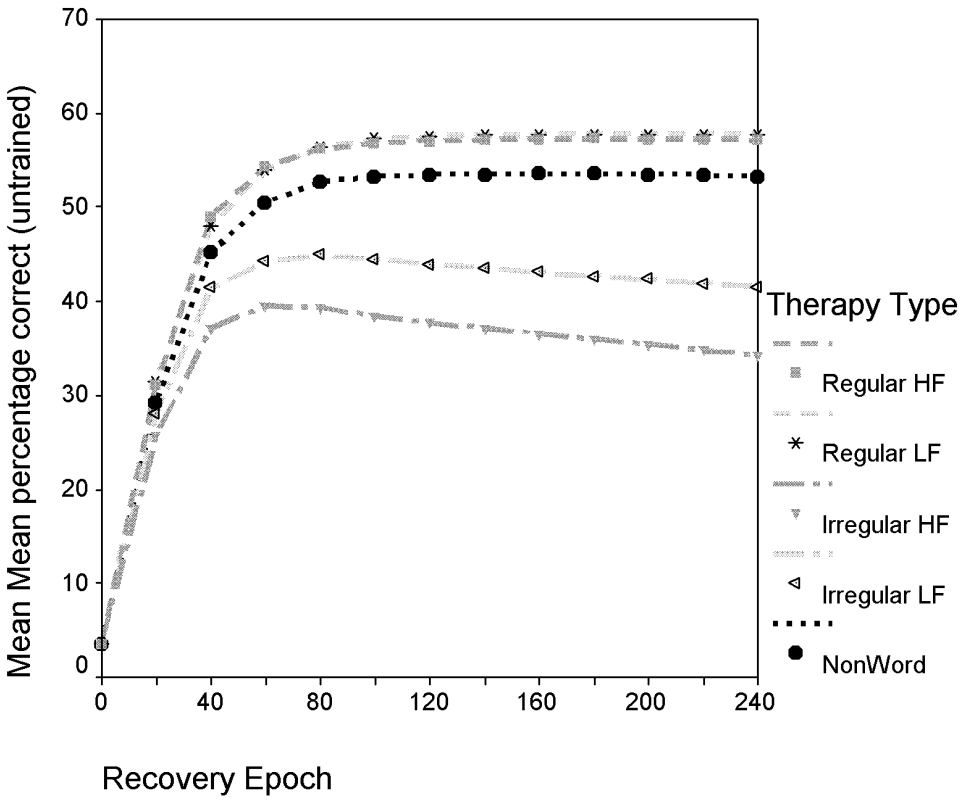

Figure 3. Recovery of reading accuracy after rehabilitation therapy.

of the network that had been trained with initial frequency weightings. Figures 4a and 4b show the results of this comparison for the high- and low-frequency irregular training sets.

While these figures are not quite identical, it is clear that all the differences in frequency (both in terms of frequency of stimuli and frequency of training set) persist in the same direction and the same magnitude even when the model is trained without any frequency weightings. Thus the only significant factors affecting the performance of the network in this simulation are consistency of training set and consistency of test stimuli set. To explore the differences in performance that result from these variables, the overall performance scores at epoch 80 were broken down on the basis of regularity of training set versus regularity of stimuli set. Figure 5 shows the results of this analysis.

This shows a very clear interaction between regularity of training set and regularity of stimuli. When training using a set of regular words, there is an enormous advantage for the network's performance on regular stimuli (nearly 60% accurate) versus its performance on irregular stimuli (less than 30% accurate). This advantage is almost completely eliminated when training on irregular words. In this case the network's accuracy rate is about 38% for irregular words and 40% for regular words. This interaction is highly significant, $F(1, 20796) = 2908$, $p < .001$. Training based on regular words results in superior overall generalisation to the original training corpus because the majority of words in English have regular spelling to sound correspondences.

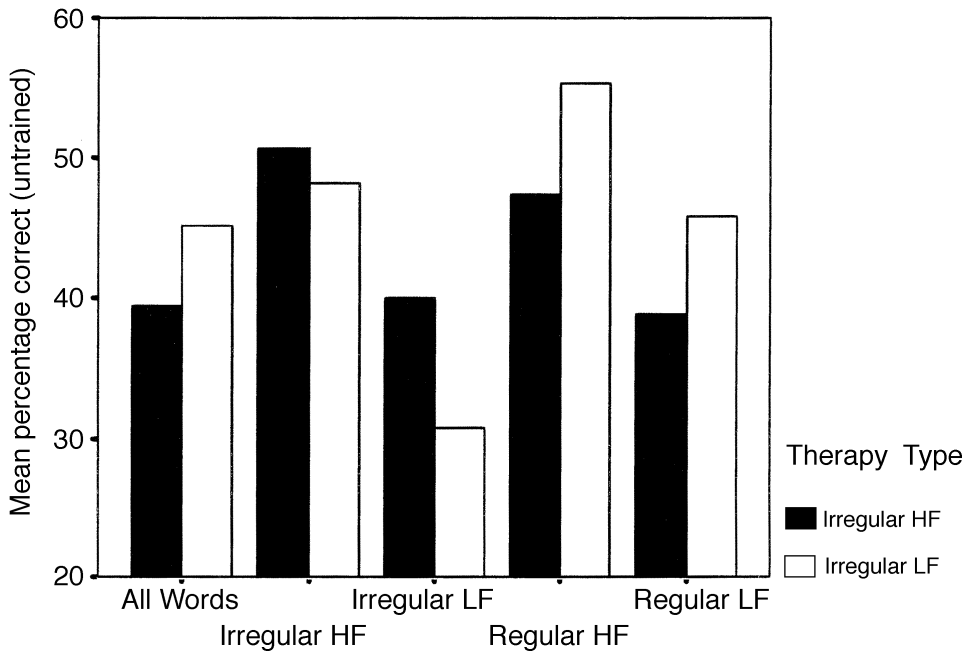

Figure 4a. Performance of the network on irregular training sets when initial training is frequency weighted.

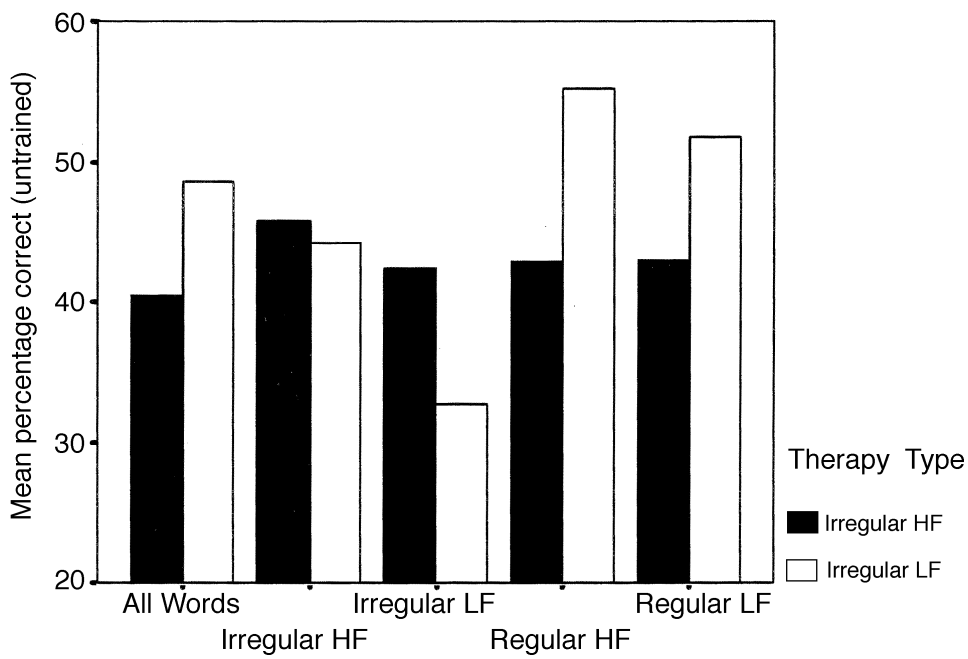

Figure 4b. Performance of the network on irregular training sets when initial training is not frequency weighted.

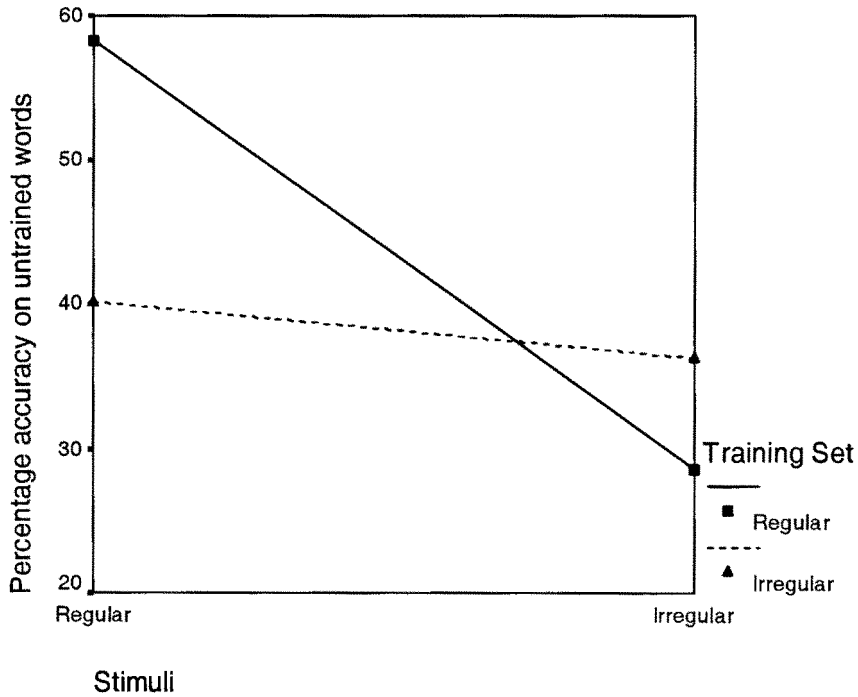

Figure 5. Network performance after 80 epochs of retraining showing the interaction between regularity of training set and regularity of stimuli set.

Discussion

The key result from this simulation is that retraining on a small set of words has considerable potential to generalise to the much larger set of untrained words. If the training set consisted of regular words then an overall accuracy of about 60% was obtained. This is a very high level of generalisation, much better than is usually found in rehabilitation studies. This is probably due in great part to the regular nature of the mappings between orthography and phonology (even for ''irregular'' words that contain many consistent elements). Good post-therapy accuracy may be partly due to the absence of any spontaneous recovery period in this simulation. This means that there is more headroom for recovery than there would be in a clinical rehabilitation study where the patient may have reached a stable, chronic level of recovered performance before rehabilitation therapy was initiated.

SIMULATION 2

This simulation was designed to explore the interaction between rehabilitation and spontaneous recovery. For these purposes rehabilitation training was defined as training using a small set of stimuli and a high learning rate (0.1), while spontaneous recovery was defined as training using the entire corpus of words but with a much lower learning rate (0.001). Online learning was used throughout the recovery and rehabilitation phase of training. (In online learning, the network updates its weights in response to each individual stimulus rather than after a whole batch of stimuli—it is computationally more intensive than batch learning but it allows for a more realistic simulation of therapeutic treatments.)

The simulation explored the effect of three variables: the spelling-to-sound consistency and frequency of the rehabilitation training set, and stage of rehabilitation therapy (early and late). In addition, a control simulation was run with no rehabilitation training to give a baseline performance for ''natural recovery'' alone. A replication of PMSP96 was trained and then damaged by removing a random 70% of the hidden units. Natural recovery was modelled by retraining the network with 40,000 word presentations selected with a frequency-weighted probability from the original training corpus. Rehabilitation therapy was modelled by interspersing the natural recovery training with presentations of the lists of therapeutic stimuli. During rehabilitation training the learning rate was raised from 0.001 to 0.1. To explore the effect of early versus late intervention, rehabilitation training was only used during one half of the period of natural recovery. Rehabilitation therapy consisted of one presentation of each of the 34 training stimuli interspersed between every 200 words from the original training corpus. Thus each training set was administered a total of 100 times during the rehabilitation phase. The rehabilitation training sets were the same as those used in simulation 1.

Results

Figure 6 shows the results of this simulation. In all cases rehabilitation resulted in a greater recovery than was achieved by unaided spontaneous recovery. Rehabilitation that occurred early in the recovery cycle was always more effective than late rehabilitation, and rehabilitation using regular items was more effective than rehabilitation using irregular items. There was also a consistent effect of frequency; items that occurred with a high frequency in the training set were more effective as therapeutic stimuli than items that occurred with a low frequency. One difference between this simulation and the previous one was the effect of training on nonwords. In Simulation 1 training on any regular training set (including nonwords) resulted in the best overall generalisation levels.

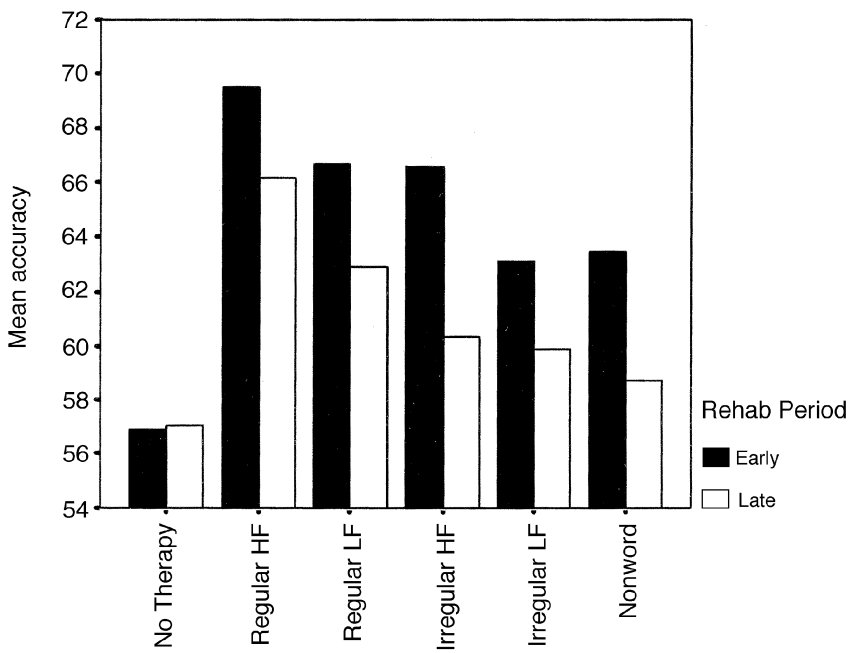

Figure 6. The effect of rehabilitation therapy as a function of therapy stimuli and timing.

In this simulation the performance after training on the regular stimuli diverged. Training on the high-frequency regular words still gave the best overall generalisation, but training on nonwords gave the worst generalisation. Further investigation revealed that this was attributable to very poor performance on irregular stimuli when training with the nonword set as opposed to training with the high-frequency regulars. To explore the effects of frequency and consistency, the improvements in scores from the no therapy condition were submitted to a 2 × 2 × 2 ANOVA where the independent variables were frequency, consistency, and timing of rehabilitation. This revealed a main effect of frequency, $F(1, 27) = 26.0$, $p < .001$, consistency, $F(1, 27) = 62.2$, $p < .001$, and timing, $F(1, 27) = 74.4$, $p < .001$. There were no significant interactions among the variables. Thus, the most effective intervention was made early in the recovery process using high-frequency regular words as therapy items.

GENERAL DISCUSSION

We have presented two simulations of reading aloud using a network architecture that replicates that used by Plaut et al (1996). The first simulation suggested that there was considerable potential for rehabilitation based on retraining using a small set of stimuli. In the best case, retraining the network on 34 high-frequency regular words resulted in the network moving from 4% to 57% accuracy on the remaining untrained corpus. The second simulation explored a more realistic model in which spontaneous recovery as well as rehabilitation training contributes to recovery. In this instance, rehabilitation training still improved performance over and above the performance improvement attributable to spontaneous recovery (69% vs 57% in the best case). In all cases, rehabilitation therapy was more effective when administered in the first half of the recovery process.

Clinical implications

Taken together, these simulations support the view that rehabilitation intervention can be useful in improving performance on both trained and untrained items. It is important to be able to show this using a model of rehabilitation, as this is notoriously challenging to demonstrate in a clinical study where it is very difficult to distinguish generalisation from spontaneous recovery: in the model we have complete control of these factors and so can address these questions more readily.

In addition to the support it gives to the view that rehabilitation can be effective, the study generated two main clinical points of interest: (1) early rehabilitation was more effective than late rehabilitation, and (2) the degree of generalisation was dependent on the choice of stimuli. This latter finding may be somewhat surprising, in that it suggests that best performance will be achieved by additional training on the items that the patient is already best at (regular words). This happens in the network for two reasons—first, because regular word reading utilises a common set of consistent spelling-to-sound correspondences, learning on one word can support learning for another. Second, as regular items greatly outnumber irregular ones in the training corpus (as they do in the language as a whole) even a small improvement for regular words in general can outweigh a larger improvement on irregulars in terms of overall reading performance.

However, before settling on an interpretation of these results it is important to be aware of the limitations of the model, which may affect our evaluation of the findings. The model we have used is a simplification of the full triangle model using only the orthographic and phonological portions (omitting semantics—word meaning). This means that it will have a tendency to be over-reliant on regularities between orthographic

and phonological representations, and it is possible that this may have exaggerated the effect of regularity. Of course, had we incorporated a semantic portion then the most obvious way of generating surface dyslexia would have been a selective lesion to this component (Plaut et al., 1996, simulation 4). In this case we would be left with only the O→P connections to support reading and our results should hold. However, if the semantic portion is only partially removed, then our simple model is not really adequate and we need to be more cautious in our interpretation. Under conditions of partial semantic impairment, therapy-driven improvements on the irregular items might involve adjustments to the damaged semantic system (or its connections to the rest of the reading system) rather than the O→P connections per se. More generally, these findings do suggest that stimulus regularity can have an effect on generalisation in domains where there is significant degree of consistency across items.

The second important finding is that therapy is most effective when administered as soon as possible after damage, irrespective of the type of stimuli used in rehabilitation. Perhaps the most plausible explanation for this is that the recovering network is at its most plastic immediately following damage so that intervention at this stage has the greatest effect. As models learn, the degree of remaining plasticity—ability to learn new information—reduces. In undamaged models this reduction in plasticity results in an age-of-acquisition effect, while in this model it emerges as a time-of-therapy effect (see Ellis & Lambon Ralph, 2000, for a fuller explanation). This result is not compromised by the simplicity of our model. Indeed it may be that the omission of a semantic portion has resulted in an understatement of the effect, as plasticity effects are generally much smaller in domains with more regular mappings (Zevin & Seidenberg, 2002). Consequently this ''time-of-therapy'' effect may be even larger in domains where the mappings were less regular than in reading (e.g., the semantics → phonology conversion underpinning speech production).

Theoretical implications

Perhaps the most important aspect of this work is that it pioneers a technique for creating a computational framework that can account for a wide range of phenomena (development, spontaneous recovery trajectories, and patterns of impaired performance) as well as making predictions concerning the effectiveness of proposed rehabilitation therapies. This technique is dependent on viewing the brain as a dynamic system that adapts both to changes in the environment (therapy) and to gross changes in computational resource (damage). Under this approach, the behaviour seen in chronic patients reflects a new dynamic equilibrium between impoverished computational resources and the demands of the environmental learning pressures (including any that arise from therapy). As we indicated in the introduction, this view contrasts markedly with both box-and-arrow and the vast majority of existing PDP models, which make the simplifying assumption that brain systems are static, with behavioural performance in the chronic phase merely reflecting the performance of the premorbid system minus those parts that have been damaged (the subtractivity assumption: Saffran, 1982). Whilst models based on this simplification have been reasonably successful in terms of diagnosis of deficits, we would suggest that progress in understanding rehabilitation will inevitably require models that adapt after damage.

The present application of this technique is relatively limited. However, we cannot see any obvious reason why this approach should not bear fruit in any cognitive domain for which an adequate computational model can be constructed (many successful models

already exist, see Introduction). A more serious limitation of the present model is that rehabilitation is limited to therapy on the same task on which we wish to improve performance. Often in rehabilitation we may want to improve performance on one task by training on another related activity. For instance in reading, one might want to concentrate on improving phonological performance by using repetition or rhyme generation tasks. As yet our model is unable to speak to these questions but we anticipate that it will be possible with a full triangle model that incorporates word meaning. This more complete model is able to simulate reading along with a number of other language activities including repetition, naming, etc. Consequently, it should be possible to investigate the efficacy of different types of therapy and their impact on other (non-treated) tasks.

Manuscript received 19 January 2005
Manuscript accepted 21 July 2005

REFERENCES

Baddeley, A. (1993). A theory of rehabilitation without a model of learning is a vehicle without an engine: A comment on Caramazza and Hillis. *Neuropsychological Rehabilitation*, *10*(3), 235–244.

Basso, A., & Marangolo, P. (2000). Cognitive neuropsychological rehabilitation: The emperor's new clothes? *Neuropsychological Rehabilitation*, *10*(3), 219–229.

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*(3), 432–446.

Buonomano, D. V., & Merzenich, M. M. (1998). Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience*, *21*, 149–186.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *The British Journal Of Psychology*, *81*(3), 361–380.

Caramazza, A., & Hillis, A. (1993). For a theory of remediation of cognitive deficits. *Neuropsychological Rehabilitation*, *3*(3), 217–234.

Coltheart, M. (2004). Brain imaging, connectionism, and cognitive neuropsychology. *Cognitive Neuropsychology*, *21*(1), 21–25.

Coltheart, M., & Byng, S. (1989). A treatment for surface dyslexia. In X. Seron & G. Deloche (Eds.), *Cognitive approaches in neuropsychological rehabilitation*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

De Partz, M. (1986). Re-education of a deep dyslexic patient: Rationale of the method and results. *Cognitive Neuropsychology*, *3*, 149–177.

Dell, G. (2004). Connectionism and cognitive neuropsychology: Comments on Harley's reflections. *Cognitive Neuropsychology*, *21*(1), 27–30.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838.

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal Of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1103–1123.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(4), 674–691.

Harley, T. (2004). Does cognitive neuropsychology have a future? *Cognitive Neuropsychology*, *21*(1), 3–16.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*(3), 491–528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662–720.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. Rumelhart, J. McClelland, & The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1: *Foundations*, pp. 282–317). Cambridge, MA: MIT Press.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences, USA*, *96*(13), 7592–7597.

Kuçera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lambon Ralph, M. A. (2004). Reconnecting cognitive neuropsychology: Commentary on Harley's ''Does cognitive neuropsychology have a future?''. *Cognitive Neuropsychology*, *21*(1), 31–35.

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 successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

McCloskey, M. (2004). Does Harley have a point? Comments on Harley's ''Does cognitive neuropsychology have a future?''. *Cognitive Neuropsychology*, *21*(1), 37–39.

Patterson, K., & Hodges, J. R. (1992). Deterioration of word meaning: Implications for reading. *Neuropsychologia*, *3*(12), 1024–1040.

Patterson, K., Seidenberg, M. S., & McClelland, J. L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neuroscience* (pp. 131–181). London: Oxford University Press.

Plaut, D. C. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and Language*, *52*(1), 25–82.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377–500.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R. et al. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235.

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216–271). Cambridge, MA: MIT Press.

Saffran, E. M. (1982). Neuropsychological approaches to the study of language. *British Journal of Psychology*, *73*, 317–337.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.

Seron, X., & Deloche, G. (Eds.). (1989). *Cognitive approaches in neuropsychological rehabilitation*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Shallice, T. (2000). Cognitive neuropsychology and rehabilitation: Is pessimism justified? *Neuropsychological Rehabilitation*, *10*(3), 209–217.

Welbourne, S. R., & Lambon Ralph, M. A. (2005). Exploring the impact of plasticity-related recovery after brain damage in a connectionist model of single-word reading. *Cognitive, Affective & Behavioral Neuroscience*, *5*(1), 77–92.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*(1), 1–29.

APPENDIX A

Words used in the rehabilitation training

Reg HF	*Reg LF*	*Irregular HF*	*Irregular LF*	*Nonword*
AIR	BREACH	BEAR	BREAST	BEED
BLACK	BROACH	BLOOD	BROOCH	BELD
BOARD	CARVE	BOUGHT	CASTE	BINK
BRIDGE	CLIFF	BREAD	CLIMB	BORT
COST	COIL	BREATH	COMB	CATH
FEEL	COUCH	DOOR	COUGH	DOLD
FOOD	DITCH	FRONT	DOST	DORE
FREE	DODGE	FULL	DOUGH	DREED
GIRL	DOLE	HEARD	GAUGE	FEAL
GREEN	GLAND	HEART	GHOUL	GODE
HAD	GLIDE	LEARN	HEARTH	GROOL
HAND	HOARSE	MEANT	HOOK	HEAN
HEAR	HOOP	MONTH	LEAPT	HEEF
HEAT	HOOT	MOST	MAUVE	HODE
LAND	LEDGE	NONE	MOULD	HOIL
LEAVE	MINCE	ONCE	MOURN	LAIL
MOUTH	MUG	POST	MOW	LOLE
MUCH	MUNCH	PROVE	PLAID	MOOP
MUST	PARE	PUSH	POLL	MUNE
NINE	PLEAT	SOME	SCARCE	NUST
OFF	PRAY	SOUL	SEIZE	PILT
PER	SAG	SOURCE	SHOVE	PLORE
ROLE	SCRIBE	SPREAD	SIEVE	PODE
SAME	SNATCH	THREAT	SOOT	PRAIN
SAW	SOUR	TOUR	SPONGE	SHEED
SENSE	SPARSE	TRUTH	STEAD	SOAD
SOUTH	STACK	TWO	STEAK	STEET
STOCK	STARCH	WHERE	SUAVE	SUFF
TOO	SWERVE	WHOM	SUEDE	SUST
TRIAL	SWOOP	WHOSE	TREAD	SWEAL
TWICE	TRANCE	WON	TROUGH	WEAT
WELL	VALE	WOOD	VASE	WOSH
WHILE	WIPE	WORLD	WOMB	WOTE
WHOLE	WISP	WOULD	YEARN	WUFF