Exploring the impact of plasticity-related recovery after brain damage in a connectionist model of single-word reading

STEPHEN R. WELBOURNE and MATTHEW A. LAMBON RALPH
University of Manchester, Manchester, England

The effect of retraining a damaged connectionist model of single-word reading was investigated with the aim of establishing whether plasticity-related changes occurring during the recovery process can contribute to our understanding of the pattern of dissociations found in brain-damaged patients. In particular, we sought to reproduce the strong frequency × consistency interactions found in surface dyslexia. A replication of Plaut, McClelland, Seidenberg, and Patterson's (1996) model of word reading was damaged and then retrained, using a standard backpropagation algorithm. Immediately after damage, there was only a small frequency × consistency interaction. Retraining the damaged model crystallized out these small differences into a strong dissociation, very similar to the pattern found in surface dyslexic patients. What is more, the percentage of regularization errors, always high in surface dyslexics, increased greatly over the retraining period, moving from under 10% to over 80% in some simulations. These results suggest that the performance patterns of brain-damaged patients can owe as much to the substantial changes in the pattern of connectivity occurring during recovery as to the original premorbid structure. This finding is discussed in relation to the traditional cognitive neuropsychological assumptions of subtractivity and transparency.

Traditionally, the science of cognitive neuropsychology has relied on evidence supplied by brain-damaged patients to make inferences about the cognitive systems subserving normal brain function. From the early work of Broca (1861) and Wernicke (1874) to the present day, a key feature of the discipline has been its reliance on information from single cases. In order to interpret this type of data, we are forced to make a number of key assumptions. These are often documented in the introductory chapters of neuropsychology textbooks (e.g., Ellis & Young, 1988; Shallice, 1988). For the purposes of this article, we are concerned primarily with two such (related) assumptions: subtractivity and transparency.

Caramazza (1984) defines transparency as the assumption that "pathological performance observed will provide a basis for discerning which component or module of the system is disrupted" (p. 10). Were this assumption to be violated on a large scale, the whole enterprise of cognitive neuropsychology would become vastly more difficult. For example, if a deficit whereby patients regularize irregular words could not be interpreted as the result of damage to a processing component in the language system, it would be difficult to see how we could progress in our understanding of brain function.

It is not necessary, however, to propose such a major dislocation between behavioral performance and cognitive structure for the transparency assumption to be threatened. A more serious threat arises from the possibility that adaptation by the damaged system may alter performance in a way that does not transparently reflect the original undamaged cognitive structure. This could occur through the growth of a new compensatory-processing module or through an existing unrelated module adapting to perform the job of the missing component. Considerations of this nature have led to the need for a further assumption known as the *subtractivity* assumption (Saffran, 1982). Ellis and Young (1988) defined it as the assumption that "the performance of a brain-injured patient is explained in terms of the normal, intact cognitive system minus those components which have been lost as a result of injury" (p. 18). As we have indicated, it is traditionally assumed that threats to the subtractivity assumption stem from the growth of new neural-tissue/cognitive modules or from the adoption of existing portions of the brain/premorbidly unrelated modules to perform new tasks. In this article, we will outline a new potential challenge to subtractivity stemming from the substantial readjustment of synaptic weights within the damaged modules themselves.

We know from clinical experience that patients' performance can change substantially in the 6 months immediately following damage, and it is possible that relearning may underpin at least a part of this recovery process. Indeed, the work of Merzenich and colleagues has demon-

strated that there is considerable plasticity in mature neural systems and that recovery of function reflects the internal reorganization of cortical representational space (for a review, see Buonomano & Merzenich, 1998) . This issue is usually avoided in neuropsychological studies, however, in which patients are tested only when their condition is thought to have stabilized. There has been a small number of studies tracking the recovery of aphasic patients (Kertesz & McCabe, 1977; Lendrem & Lincon, 1985; Wade, Hewer, David, & Enderby, 1986). Unfortunately, all of these are neurologically, rather then neuropsychologically, motivated, and they do not contain the kind of detailed neuropsychological tests that allow one to interpret the patterns of performance in terms of changes to a cognitive framework. Although lacking in neuropsychological detail, these studies do demonstrate that there is considerable improvement in overall aphasic language performance in the period immediately following onset. Lendrem and Lincon reported that this *spontaneous* recovery occurred irrespective of therapeutic intervention and was greatest during the first 3 months postonset. This recovery was so substantial that almost half of the patients recovered to the point where they could no longer be classified as aphasic.

Spontaneous recovery could be attributed to a number of mechanisms: It could simply be the result of physiological recovery, where the activity of neural substrates is suppressed (perhaps due to transient oedema), so the improvement in performance represents these partially damaged parts coming back "on line"; conversely, recovery could be attributed to the recruitment of other brain structures; finally, spontaneous recovery could be a relearning process in which the surviving neurones in the damaged systems readjust their connection weights to provide the best possible performance that can be achieved from the impoverished computational resources. In reality, all three mechanisms may play a part in the improvement of performance after brain damage. For the purposes of this article, however, we will focus solely on the possibility that spontaneous recovery is essentially a relearning process.

This study was based on a parallel distributed processing (PDP) computational simulation of single-word reading and focused specifically on how the performance of the damaged model changed during a period of recovery. PDP models of cognitive function are particularly appropriate for this kind of study, since they provide a computational analysis that bridges the gap between neuroscience and behavior. Because PDP models learn, they allow a direct, formal exploration of both behavioral and computational aspects of plasticity.

Because we were focusing on the process of recovery, it would have been possible to model one of any number of cognitive processes .We selected single-word reading, since there was already a wealth of cognitive neuropsychological data in this domain, as well as some very successful connectionist models that could serve as the starting point for our investigations. Despite the success

of connectionist models of reading, there has been very little work to date on recovery following damage, and 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 generalize to the rest. More recently, Plaut (1996) used a version of his deep dyslexia model (see 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 generalization than did typical ones. (See Kiran & Thompson, 2003, for empirical support of this prediction.)

These models differ from the present simulations in two critical respects. First, previous models (with the exception of Simulation 3 from Plaut, 1996) adopt the *traditional* premise (stemming from the subtractivity assumption) 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 the modeling of rehabilitation and, in particular, the ability of models to generalize from retraining on a reduced training corpus to the remaining items. By contrast, we suggest that the lesioned model may be analogous to a patient immediately after brain damage has occurred and that, rather than modeling rehabilitation, we are modeling spontaneous recovery through a retraining process. Accordingly, to test how a trained set of items can generalize to an untrained set (to mimic rehabilitation interventions), we have no need to restrict the training corpus; instead, we assume that the environmental factors contributing to spontaneous recovery are similar to those contributing to original learning. (We acknowledge that the overall level of exposure to reading is likely to be reduced postmorbidly, but we expect that the relative influence of individual words will not be significantly altered.) Thus, spontaneous recovery can be modeled by continuing training with the full set after the damage has been inflicted. This approach allows us to focus on how the pattern of performance changes through the retraining process and on whether performance after retraining is a better match to brain-damaged patients than performance immediately after damage is. The second novel feature of this recovery simulation is that it models the process of translating orthography to phonology directly, rather than via semantic representations (Plaut, 1996).

A good computational model of single-word reading should be able to reproduce a number of features of reading that are found in human populations. In addition to learning the pronunciations of the words in the training set, a good model should also be able to generalize its learning to regular nonwords; it should reproduce a strong frequency $\times$ consistency interaction in word-

naming performance, and it should be able to reproduce some of the patterns of errors found in brain-damaged patients. The first reading model that came close to any of these objectives was Seidenberg and McClelland's (1989) model of word reading—hereafter, known as *SM89*. SM89 is a restricted implementation of the larger *triangle* model. Whereas 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 2,820/2,897 monosyllabic words (97.3%). What is more, the model displayed the same kind of frequency $\times$ 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 score than low-frequency consistent words did. However, the model performed at a level considerably below that for humans when generalizing to nonwords; Besner, Twilley, McCann, and 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.

Patterson, Seidenberg, and McClelland (1989) investigated the possibility that lesioning SM89 would produce symptoms of surface dyslexia. Although there was some match between the performance of the model under damage and that found in surface dyslexics, it was not very compelling. In particular, the model did not display a sufficiently large consistency $\times$ regularity interaction when damaged. Neither did it produce enough regularization errors to make a convincing case that it was modeling surface dyslexia (Patterson, Seidenberg, & McClelland, 1989; Plaut, McClelland, Seidenberg, & Patterson, 1996).
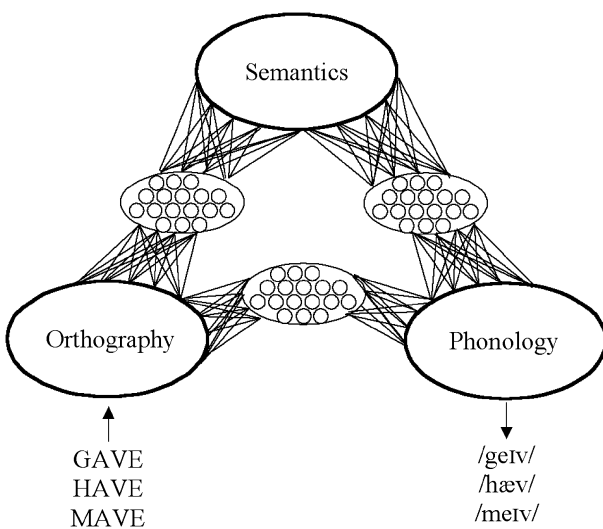

Figure 1. The triangle model from Seidenberg and McClelland (1989).

Plaut et al. (1996) returned to the issue of modeling single-word reading. 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 utilized, which suffered from what they termed the *dispersion* problem. They proposed a new model based on a more condensed representational structure, which greatly reduced this problem.

This model (Simulation 2 in Plaut et al., 1996)—hereafter PMSP96—was trained on a set of 2,972 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. Unlike SM89, it also performed at human levels in nonword reading; it could correctly read 97.7% of the consistent nonwords from Glushko (1979). In addition, the model displayed the required frequency $\times$ consistency interaction so that it fulfilled all the basic requirements for a good model of word reading. When the model was lesioned, however, it did not perform in a way that mimicked surface dyslexia. 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 regularization errors, although 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 modeled 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 modeled by training the network in the presence of a semantic input and then removing that semantic contribution.

We took PMSP96 (Simulation 2) as our starting point. We speculated that Plaut et al. (1996) might have given up on the single-route model too soon and that it is possible to model surface dyslexia within a single route if the effect of retraining is considered. This does not mean that we are arguing against the idea that surface dyslexia can arise from a selective semantic impairment; indeed, we think it very likely that it does—the majority of classical surface dyslexics reported in the literature are those with semantic impairment, most commonly in the context of semantic dementia or profound temporal lobe damage (Bub, Cancelliere, & Kertesz, 1985; Patterson & Hodges, 1992). It may be that the symptoms of surface dyslexia, however, can arise from more than one kind of damage. In any case, the main thrust of this investigation is to explore the process of recovery through a tangible example of retraining a single-route reading model. This study demonstrates that retraining a replication of PMSP96 can result in a significantly better match to surface dyslexic symptoms than can be obtained, in the same network, immediately after damage. The possibility of internal reorganization, leading to substantial changes in overt behavior, is in itself interesting and may also have wider implications requiring a refining of the

key cognitive neuropsychological assumptions of subtractivity and transparency.

SIMULATION 1
Traditional Account of Surface Dyslexia, Using a Replication of PMSP96

The first simulation provided a baseline for further work. Essentially, it was a replication of the feedforward simulation from PMSP96, which was then damaged in varying degrees and locations to assess how the network would perform immediately after damage with no retraining. This differs slightly from the results in Plaut et al. (1996), since they only reported the effects of damage to a more complex attractor network. In that respect, this simulation is probably closer to the work by Patterson, Seidenberg, and McClelland (1989) on the SM89 network.

Network Architecture and Learning Algorithm

The architecture, training, and representations used in this simulation were chosen to be as similar as possible to those used by Plaut et al. (1996, Simulation 2).[1] Each of these key features is summarized 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. Cross-entropy was used as the error function, as in PMSP96. The learning rate for the network was set to 0.05, and the momentum was 0.9.
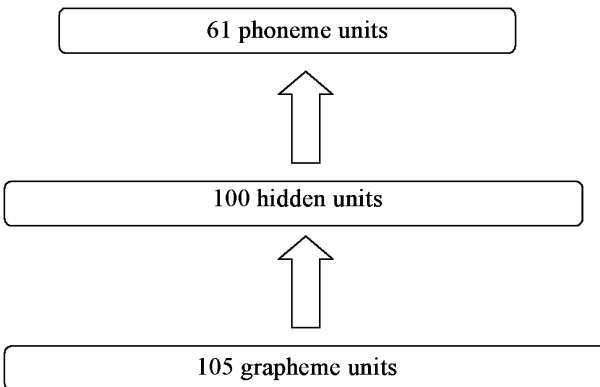

Figure 2. Network architecture.

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* (Jacobs, 1988). 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 did PMSP96, designed to minimize the dispersion problem. 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 phonological onset and coda are further divided into groups of mutually exclusive phonemes, so that when the unit activations are read off, 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, the output is read in the order that they appear in the table. The only exception to this occurs for the phoneme 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, they are taken in the order sp, unless the ps unit is active, in which case the order is reversed.

Training Procedure

The network was trained using full batches with the same corpus of 2,998 monosyllabic words used in PMSP96. The frequency (Kučera & 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's being presented for training; however, it has the considerable advantage that every word can still be presented once every epoch, thus 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 was tested every 10 epochs throughout the training period. Five sets of test stimuli were used in this testing process: (1) high-frequency regular words ($n = 24$) (*best, big, came . . .*), (2) high-frequency irregular words ($n = 24$) (*are, both, break . . .*), (3) low-frequency regular words ($n = 24$) (*beam, broke, bus . . .*), (4) low-frequency irregular words ($n = 24$) (*bowl, broad, bush . . .*), and (5) regular

Table 1
Orthographic and Phonological Representations

Part	Units
	Orthographic
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 UY
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 SL SS TCH TH TS TT ZZ U E ES ED
	Phonological
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
Coda	r l m n N b g d ps ks ts sec z f v p k t S Z T D C j

nonwords (*n* = 43) (*beed*, *beld*, *bink* . . .). The regular and irregular words were taken from Taraban and Mc-Clelland (1987) and were matched across groups for frequency. The regular nonwords were taken from Glushko (1979) and were created by changing the onset of an existing regular word. All of these stimuli are the same as those used in PMSP96, so that it is possible to make a direct comparison of results.

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.

In addition to the performance on the five sets of test stimuli, the percentage of regularization errors made by the network on the two irregular stimuli sets was also recorded. This was calculated by matching the actual output of the network against a list of possible regularizations for each word. For most words, this list consisted of just one possible regularization, but some words (e.g., *flood* or *lose*) can be regularized in two different ways. Again, the list of pronunciations treated as regularizations was taken from those used in PMSP96.

Developmental Progress of PMSP Replication

Figure 3 shows the developmental progress of the network averaged across the 10 training runs. As we would expect, learning was slowest for the low-frequency irregular words, with the high-frequency regular words
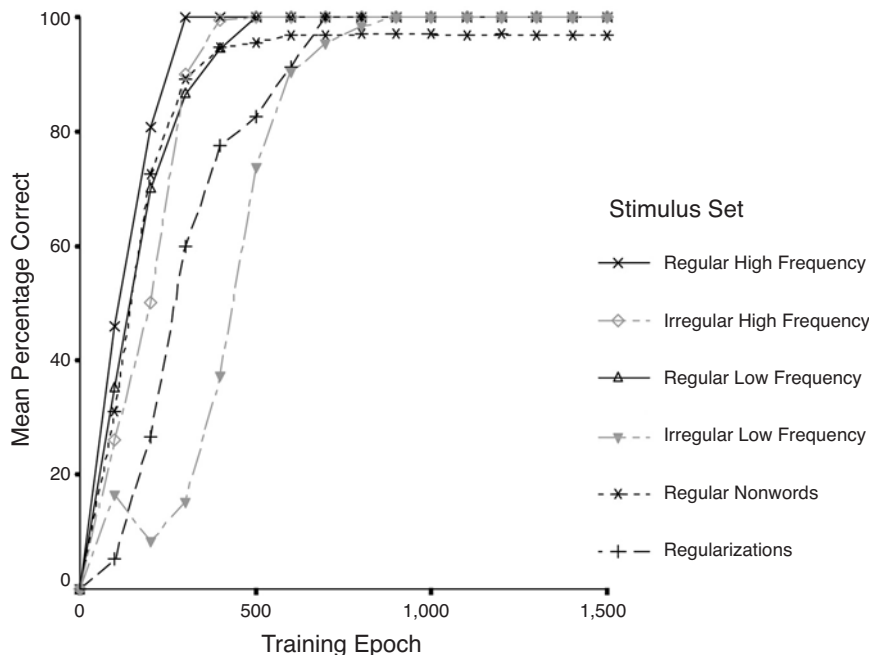

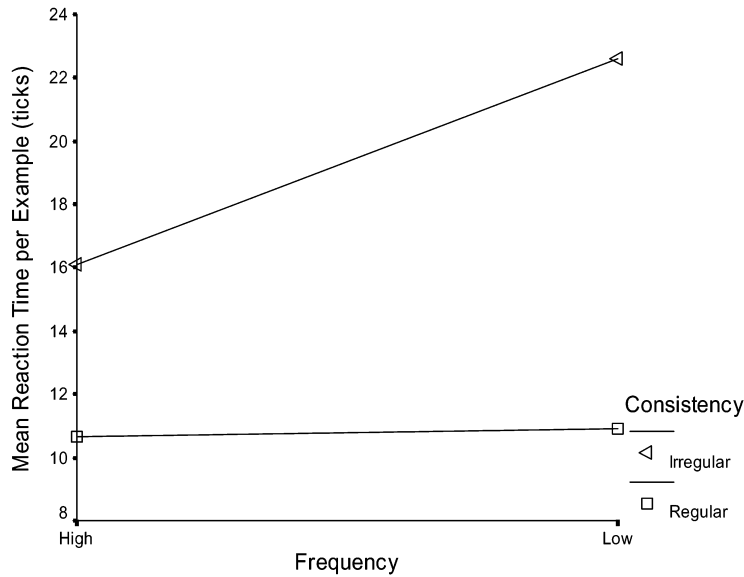**Figure 3. Network performance as a function of training epoch.**

Figure 4. Mean reaction time as a function of frequency and consistency.

being the fastest. By Epoch 1,000, the network was performing optimally for all of the stimulus 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 1 of the 10 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 performed at 99.97% accuracy, which is 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 correctly read 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).

It is important to verify that this model could replicate the standard frequency × consistency interaction found in the naming latencies of normal human populations (e.g., Seidenberg, 1985; Seidenberg, Waters, Barnes, & Tanenhaus, 1984). Strictly speaking, our stimuli design allows for testing of frequency × regularity, rather than frequency × consistency, interactions, but following Plaut et al. (1996), we prefer to describe them in terms of consistency, since, in our model the effects arise as a result of the network's detecting consistencies between the regular stimuli, rather than from the application of a set of grapheme-to-phoneme conversion rules.

In order to obtain naming latencies from our feedforward network, we adopted a technique from Cohen, Dunbar, and McClelland (1990). We time averaged the input to each unit (except the input units) across 100 ar-

bitrary time units (ticks), using a time-averaging proportion of .1, and added time-dependent Gaussian noise with an *SD* of 0.1. Using weights from Epoch 1,000 of the training, we ran each of the 10 networks for 10 trials on each example, taking the response time as the number of ticks required for the output to reach criteria. A 2 × 2 analysis of variance was applied to these reaction times, where frequency and consistency were treated as between-group variables. This confirmed that there was indeed a significant frequency × consistency interaction [$F(1,396) = 968.2, p < .001$]. In addition, there were significant main effects of both frequency [$F(1,396) = 1144.4, p < .001$] and consistency [$F(1,396) = 7351.3, p < .001$]. Figure 4 shows this interaction in detail. Clearly, the effect of frequency was almost completely modulated by consistency. For irregular words, low frequencies resulted in a much higher RT, but for regular words there was almost no effect of frequency. This is consistent with the standard effect found in human reading latencies and with the results found for PMSP96.

Results From Simple Lesioning of the PMSP96 Simulation: Traditional Account of Surface Dyslexia

To investigate how well the model accounted for surface dyslexia under a traditional theoretical framework, the network was damaged by lesioning each of the three possible locations (input–hidden links, hidden units, and hidden–output links). The proportion of units/links lesioned varied from 0 to .8 in steps of .02. For each combination of lesion severity and location, 10 trial lesions were performed, and this process was repeated for each of the 10 trial networks, so that each data point repre-
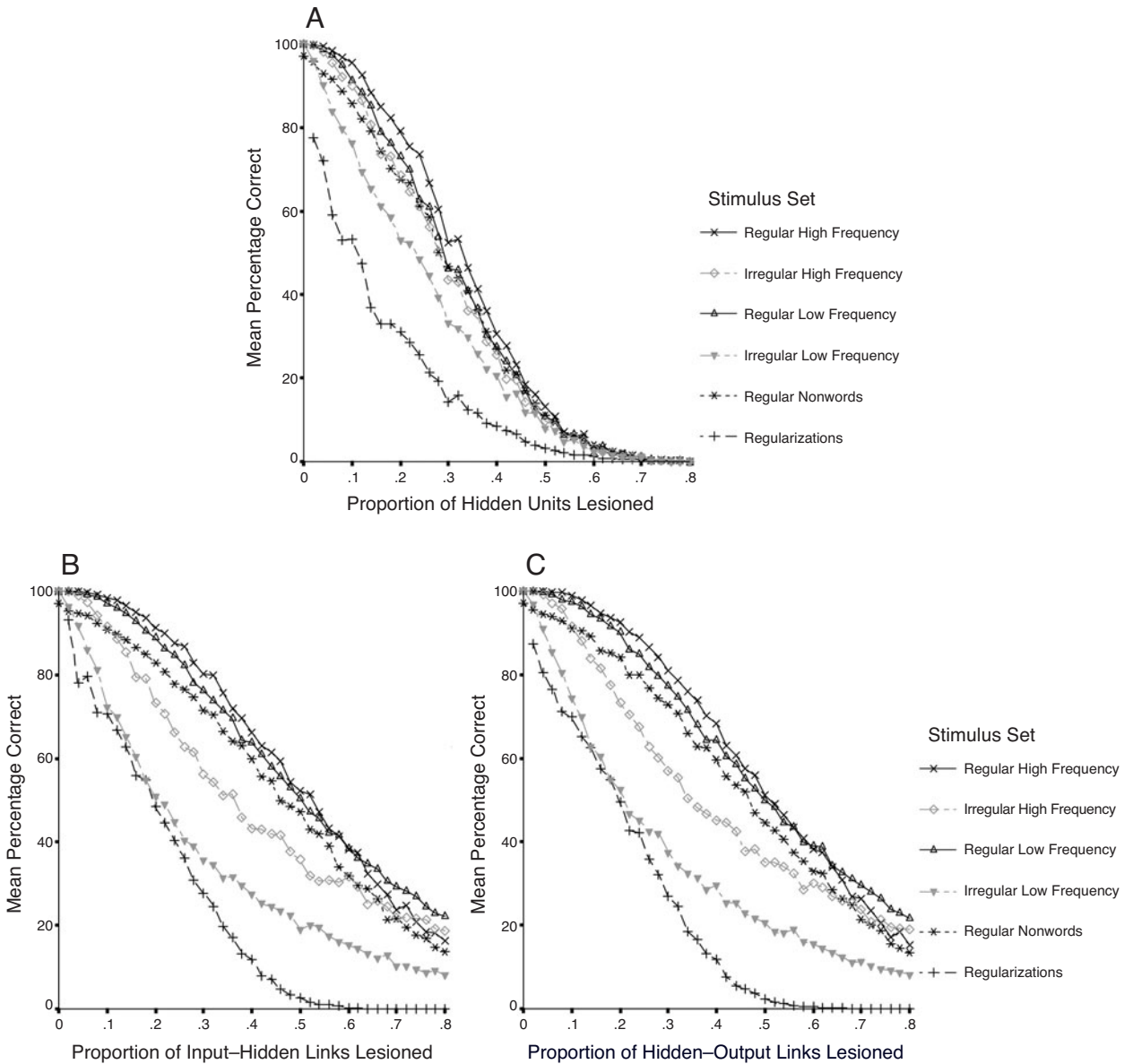
Figure 5. Network performance following lesions to (A) hidden units, (B) input–hidden links, and (C) hidden–output links.

sents the average of 100 trials. Figures 5A–5C show how the networks' performance was affected by these lesions for each of the three possible lesion locations.

For all of the locations, there was a similar story. As we would expect, damage resulted in impaired performance on all of the stimuli types. High-frequency regular words performed best, and low-frequency irregular words performed worst, with the other stimuli sets falling in between. The difference in performance between high-frequency regulars and low-frequency irregulars was greater for damage to the links (either input–hidden or hidden–output) than for damage to the hidden units. Also, damage to the hidden units resulted in a greater

overall performance degradation. This is what one would expect, since damage to a hidden unit is effectively the same as damaging all the incoming and outgoing links to that unit, so that damaging 50% of hidden units will effectively remove 50% of the input-to-hidden links, as well as 50% of the hidden-to-output links.

These results are not very convincing as a model of surface dyslexia. Although there was a small frequency × consistency interaction, it was not anywhere nearly as large as that found in severe surface dyslexic patients. What is more, the number of regularization errors was much less than that found in surface dyslexics. This is in line with the results from other modeling studies taking

the same traditional view of surface dyslexia (Patterson et al., 1989; Plaut et al., 1996), and it confirms that this model does not give a good account of surface dyslexia when viewed from this standpoint.

SIMULATION 2
Effect of Retraining After Damage: An Alternative Account of Surface Dyslexia

This simulation used the same network architecture and training parameters as those in Simulation 1. The 10 fully trained networks were damaged with varying degrees of severity and then were retrained for 600 epochs,

using the same training corpus as that in the original training, weighted by frequency. We explored the effect of damage to all three possible locations, but for the sake of brevity, we will focus mainly on recovery following damage to the hidden units. This type of damage produced the smallest frequency $\times$ consistency interaction in Simulation 1 and thus presents the biggest challenge in trying to model surface dyslexia. Figure 6A shows the performance of the model as it recovered from a 60% lesion to the hidden units. Perhaps the most striking features of this were the speed and extent of recovery. Immediately after damage, the network performed at less than 10% accuracy for all the stimulus types; yet after
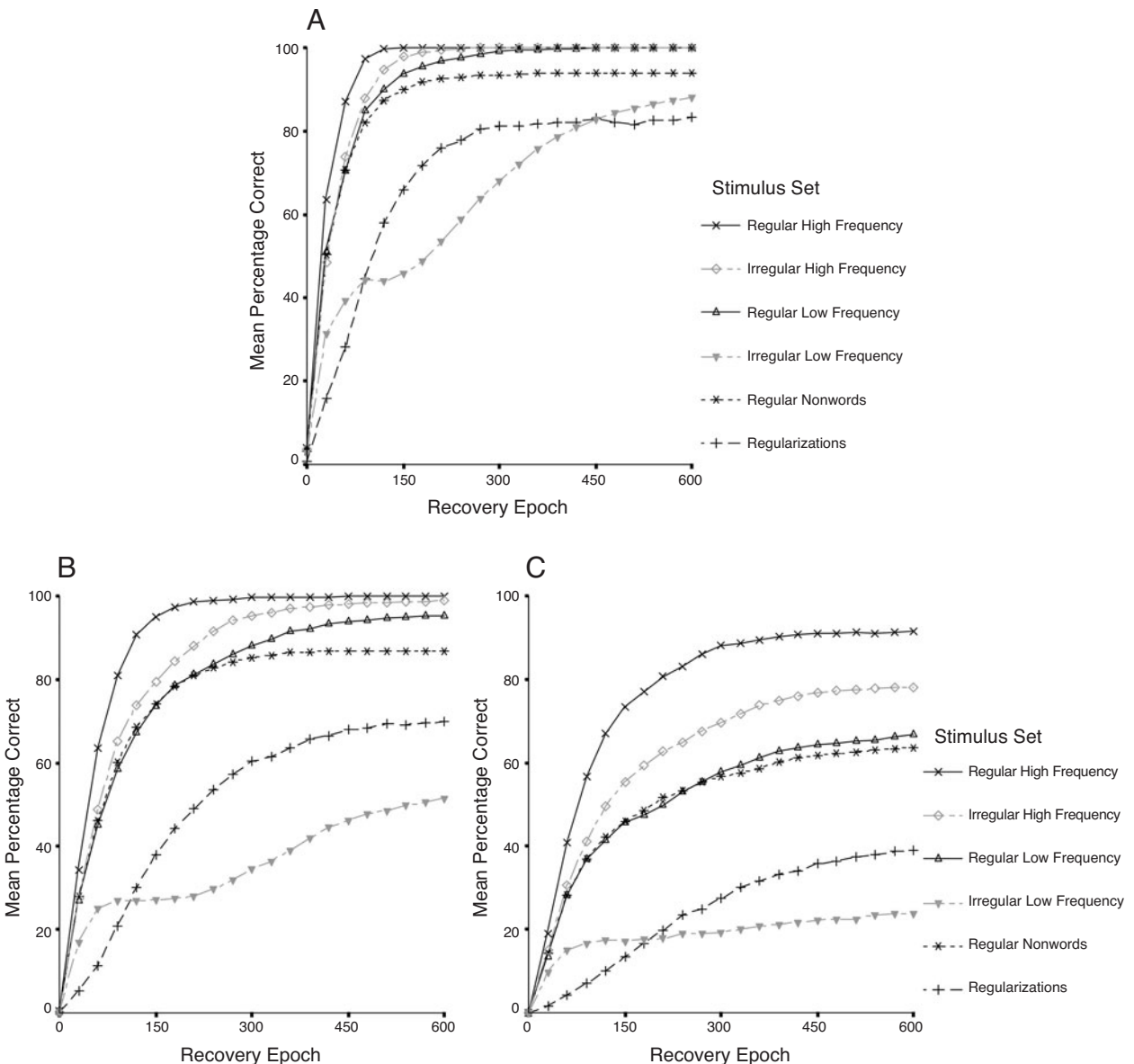

Figure 6. Time course of recovery following (A) a 60% lesion to hidden units, (B) a 75% lesion to hidden units, and (C) an 85% lesion to hidden units.
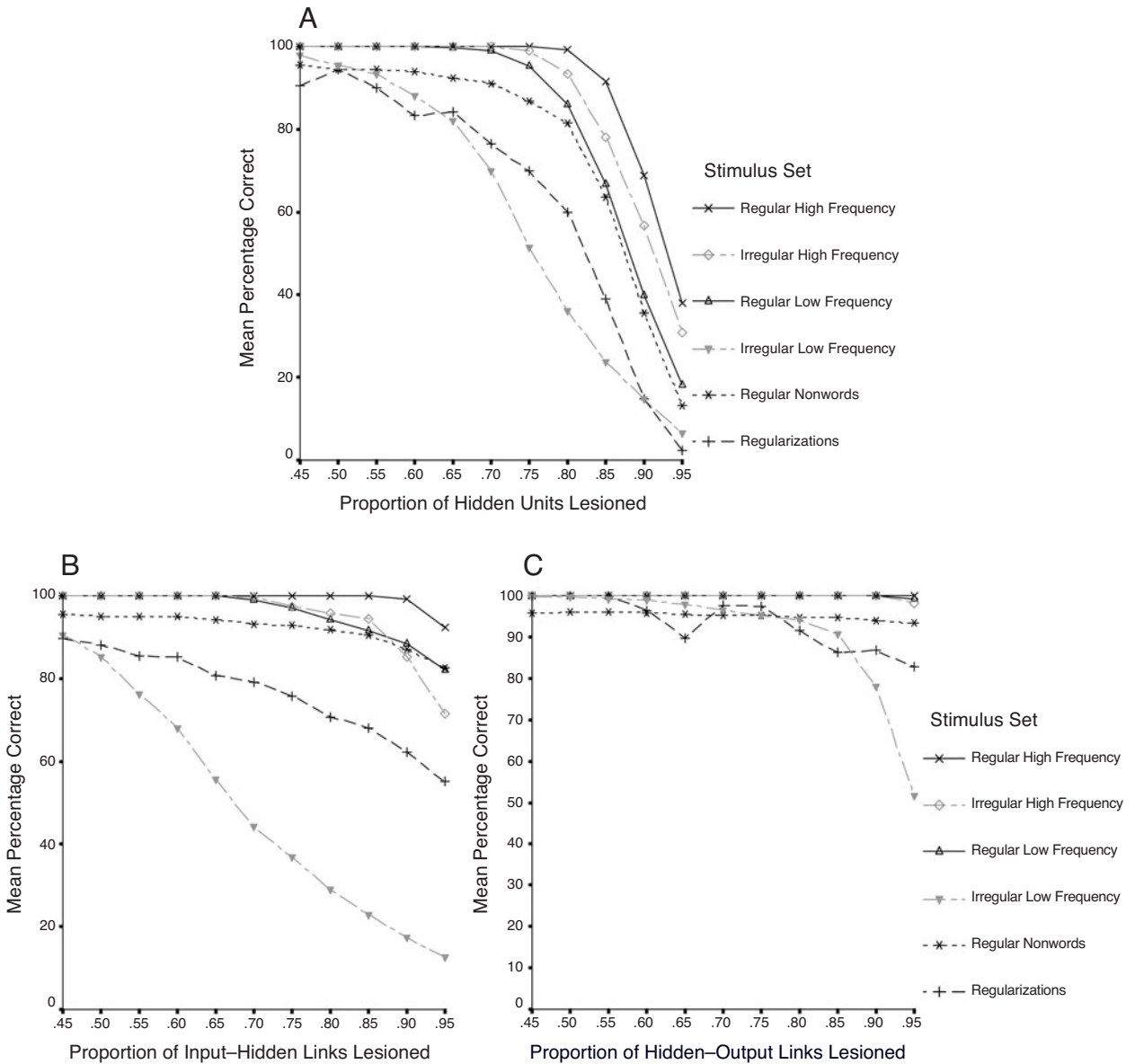
Figure 7. (A) Performance after 600 epochs of retraining following varying severities of damage to (A) hidden units, (B) input–hidden links, and (C) hidden–output links.

only 90 epochs of retraining, it had recovered to 80% correct on all except the low-frequency irregular words. In addition, the pattern of performance over the different stimuli was a very good fit to the pattern that we would expect to find in surface dyslexic patients. From Epochs 100–600, regularization errors consistently made up more than 50% of the errors to irregular words. While performance on most stimulus types was between 85% and 100%, performance on low-frequency irregular words started at 45% and only slowly improved, reaching 88% by Epoch 600. At any time from Epoch 100 onward, the network would clearly be classified as surface dyslexic,

with the severity of the diagnosis decreasing with additional retraining.

Figure 6B shows the effect of a retraining after a more severe (75%) lesion to the hidden units. Again, there was a very rapid recovery for all except the low-frequency irregular words. However, in this case, the network could not quite achieve such a complete recovery. By Epoch 600, performance on the high-frequency stimuli was close to 100%; it reached 95% on the low-frequency regular words and 85% for the nonwords—slightly impaired as compared with normal human performance. However, only about 50% of the low-frequency irregular words were correctly

pronounced. Just under 70% of these errors are the result of regularizations; this is perhaps slightly less than we might find in the purest cases of surface dyslexia, but it is still clearly characteristic of the disorder.

Figure 6C shows the effect of retraining after a very severe (85%) lesion to the hidden units. At this level of lesion, performance immediately postdamage was at zero, and yet after retraining, performance improved enormously. Although all types of stimuli were impaired to some degree, there was still a very large frequency × consistency interaction; nonwords and low-frequency regular words were pronounced correctly about 60% of the time, performance on high-frequency regular words was the best at 92% correct; reading of high-frequency irregular words fell somewhere in between at 78% correct; but by far the worst performance was on the low-frequency irregular words, which were pronounced correctly in only 24% of cases. Of the errors made on irregular words, 39% were regularizations. So even in this extreme case, some of the typical surface dyslexic patterns were present, albeit in an impure form.

So far, we have looked only at the effect of retraining following damage to hidden units. To compare the effect of retraining following lesions at different locations, the network was damaged with varying severity and was retrained for 600 epochs. Figures 7A, 7B, and 7C show the profiles of performance as a function of lesion severity for the three possible lesion locations. This revealed two main trends. First, the overall performance was dependent on lesion location. Damage to the hidden–output links gave the least overall impairment, with damage to the hidden units resulting in the most serious performance degradation. In addition, the percentage of regu-

larization errors was highest for lesions to the hidden–output links, with regularization errors consistently greater than 90% for damage levels up to 80%, although it should be noted that, in these cases, performance on all the words is very good, so the network is performing at near-normal levels. For damage to hidden units or input-to-hidden links, this level of regularizations was maintained only for damage levels up to 55% and 45%, respectively. Second, the magnitude of the frequency × consistency interaction was also modulated by lesion location. A rough measure of the size of this interaction would be the level of impairment of low-frequency irregular words at the last point for which performance on the high-frequency irregular words is still 100%. For lesions to the hidden units or hidden–output links, this measure yields roughly a 30% error rate (at 70% and 90% damage, respectively), but for lesions to the input–hidden links, it gives an error rate of about 55% (at 70% damage), indicating that this damage location gives the best fit to *pure* surface dyslexia. This is illustrated more clearly in Figure 8, which shows the time course of recovery following an 85% lesion to the links between input and hidden units.

INTERIM SUMMARY FOR SIMULATIONS 1 AND 2

The principal aim of this study was to investigate the effect of retraining a PDP model of word reading, to see whether changes occurring during retraining had the potential to explain subsequent dissociations in performance patterns. In Simulation 1, we replicated the feedforward simulation from PMSP96 and confirmed that damage to this network does not produce patterns of per-
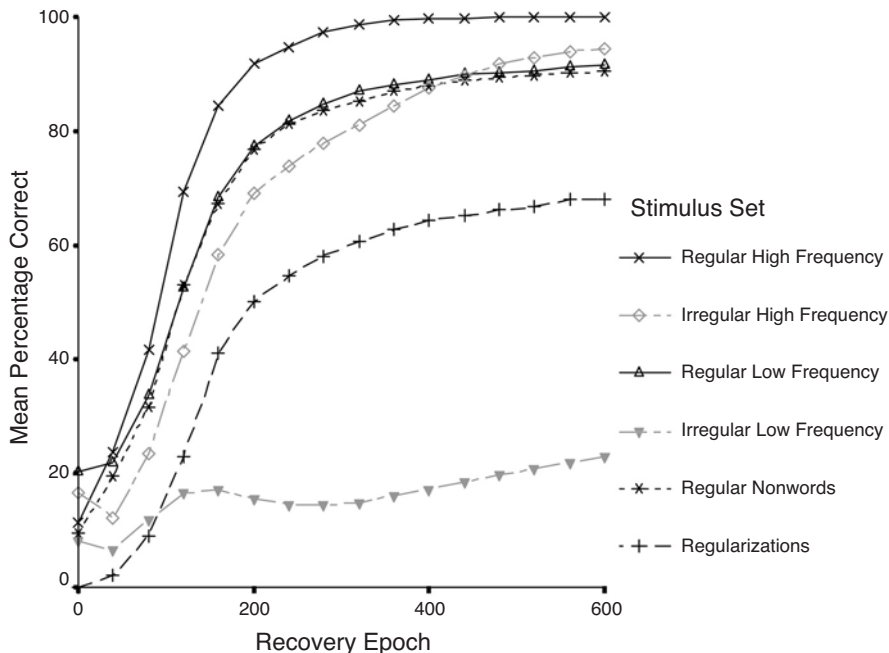

Figure 8. Time course of recovery following an 85% lesion to input–hidden links.

formance that can be characterized as surface dyslexia. In particular, the frequency × consistency interaction was too small, and there were too few regularization errors. However, in Simulation 2, we demonstrated that retraining the network after damage *crystallizes* out these small dissociations and increases the number of regularization errors dramatically. In this case, the pattern of performance was strongly reminiscent of the pattern found in surface dyslexic patients, with a strong frequency × consistency interaction and a large number of regularization errors. As was discussed earlier, these results pose a potential challenge to the standard neuropsychological assumptions of subtractivity and transparency. Clearly, in our model, the dissociation between low-frequency irregular word and nonword reading owes at least as much to the reorganization of connection weights posttrauma as to the original premorbid architecture.

SIMULATION 3
Modeling Progressive Aphasia

So far, we have concentrated on modeling patient performance with a major lesion followed by some retraining. Although this is a reasonable approximation of what happens for patients following a stroke or a head injury, it is not accurate for patients with a progressive disease in which the damage is incremental over a long period of time. Interestingly, almost all of the surface dyslexic cases reported in the literature are associated with a progressive, neurodegenerative condition (semantic dementia), and it seems that this form of damage produces the purest cases—for example, Patient K.T. (McCarthy & Warrington, 1986). This raises the question as to whether it is appropriate to compare data from a model that has

simulated a single acute lesion followed by recovery with data from patients who suffer from a progressive disease. The development of the disease may have a profound effect on the observed patterns of deficits. Indeed, the principal theme of this study is that the process of recovery can be critical in the formation of behavioral dissociations. It seems reasonable, therefore, to postulate that the process of continual, incremental damage, coupled with a constant environmental pressure to perform accurately, might result in even more severe behavioral dissociations than those obtained from a single massive episode of damage followed by retraining.

To test this hypothesis, the 10 fully trained networks were subjected to repeated episodes of damage (1% of the input–hidden links), followed by 20 epochs of retraining. Figures 9A and 9B show the results of this simulation. Figure 9A shows the performance of the network as a function of degeneration epoch, and Figure 9B shows the same performance as a function of lesion severity (see Figure 7B for the comparable network with an *acute* lesion followed by retraining). Two features stand out immediately. First the frequency × consistency interaction was more pronounced than in any of the previous simulations. At the point at which 70% of the links were lesioned, the low-frequency irregular words were read with only 25% accuracy, whereas the high-frequency words were read with 96% accuracy. The comparable figures for the acute case were 45% and 100%, respectively. This would seem to support the hypothesis that the more extreme dissociations of performance found in semantic dementia patients such as K.T. can be explained, at least in part, by the incremental nature of the damage. It should be noted, at this point, that although this simulation is a very good match to patients such as K.T. in
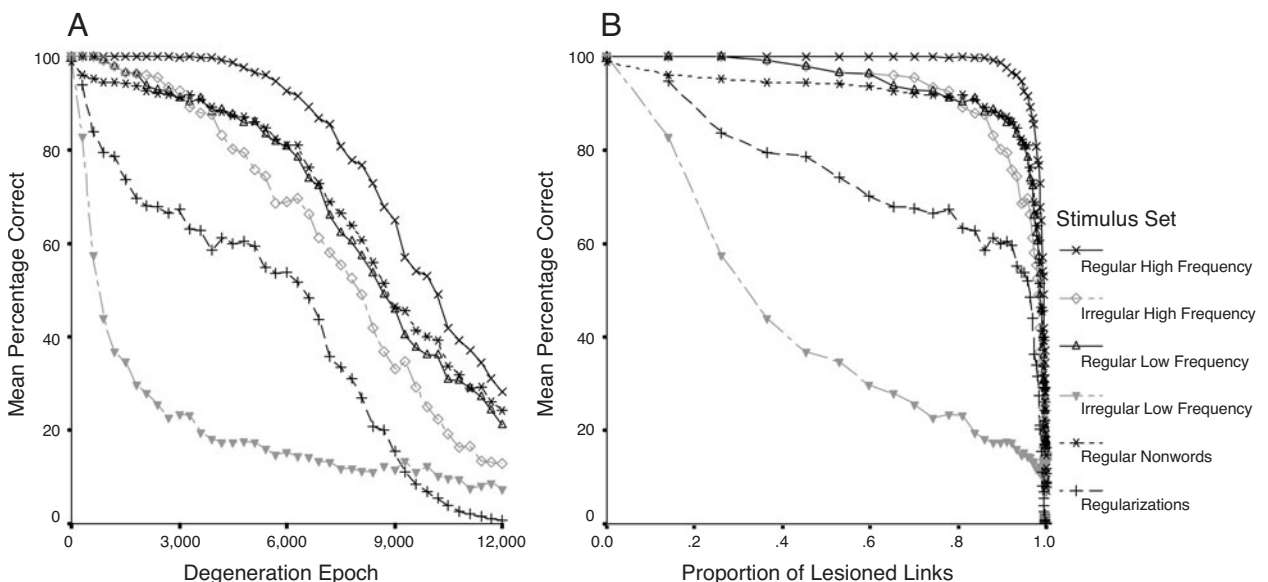

Figure 9. The effect of progressive lesions to the input–hidden links (A) in an environment of continual training and (B) as a function of lesion severity.

terms of the magnitude of the frequency × consistency interaction, it does not produce quite the level of regularization errors that one might expect from very pure surface dyslexics. The network is producing regularizations in the region of 70%–80%, whereas pure surface dyslexics such as K.T. might be expected to regularize nearer 90% of the irregular words that they read incorrectly.

In addition to the magnitude of the frequency × consistency interaction, it is interesting to note that this strong interaction was maintained through a considerable fraction of the degeneration time course. This observation may explain why so many of the reported cases of surface dyslexia come from patients with semantic dementia, who also maintain a very strong frequency × consistency interaction throughout a large fraction of the course of their disease (Patterson & Hodges, 1992). In contrast, the simulations suggest that to obtain a similar-sized interaction from a single acute insult requires the damage to be in a relatively confined range of severity. These results predict that nonprogressive cases of surface dyslexia will tend to have less extreme frequency × consistency interactions than those found in the progressive cases.

SIMULATION 4
Are Developmental Disorders Different From Cases of Adult Brain Damage?

An issue that has sparked considerable controversy within the field of cognitive neuropsychology is whether developmental disorders can be considered in the same framework as disorders arising from adult brain damage. Many researchers in the field see no fundamental differ-

ence between these two classes of disorder, arguing that the data are behaviorally indistinguishable (see Baron-Cohen, 1998; Temple, 1997). Others have argued strongly that there is a qualitative difference between the two cases, arising from the interaction of the developmental process and the deficit (see Bishop, 1997; Karmiloff-Smith, 1997). This controversy has recently been examined in the domain of connectionist modeling in work by Thomas and Karmiloff-Smith (2002). They used a simple model of past tense verb production and contrasted the effect of damage pre- and postdevelopment. They found that the behavioral dissociations between regular and irregular items were magnified in the case in which damage was inflicted prior to training. They concluded that this supported the idea that developmental disorders were essentially different from those found in adults. However, in light of our previous simulations, another interpretation of these results is possible. Perhaps the *developmental* simulations provide the best results for both developmental and adult disorders. We have suggested that our understanding of some adult disorders may be improved by consideration of the effects of postdamage retraining. This retraining period might well allow for the same interaction between deficit and development that occurs in the developmental case. On the other hand, it may be that reduced plasticity in the adult model (Ellis & Lambon Ralph, 2000) will reduce the model's ability to adapt, so that the difference between developmental and adult disorders is maintained even when the effect of retraining after damage is considered.

To investigate this, we compared the effects of two separate training regimes: In the first, *acquired dyslexia*
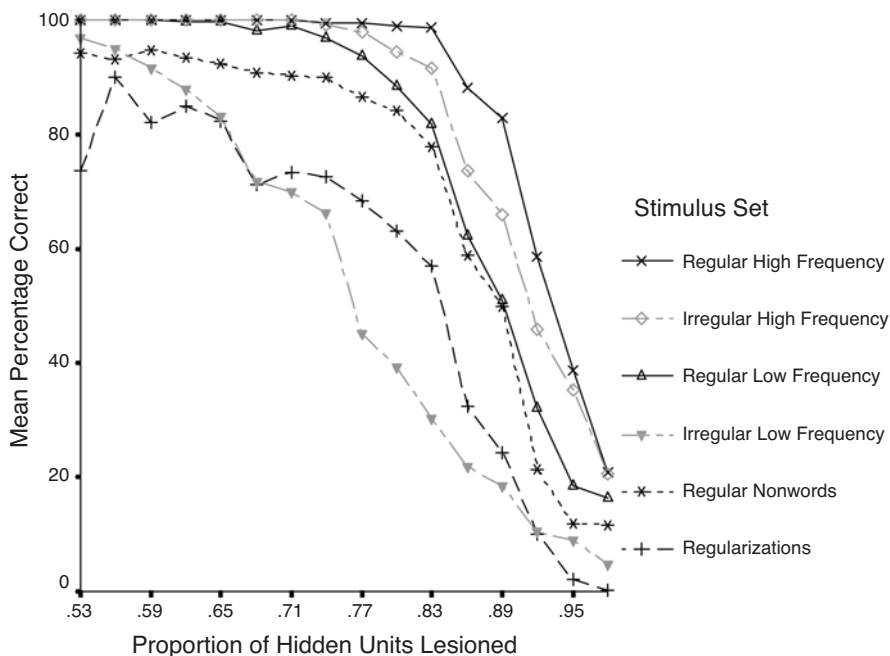

Figure 10. Acquired surface dyslexia curve, which shows the performance of the network after the following sequence: train for 1,000 epochs, damage, train for an additional 800 epochs.

regime, the model was trained for 1,000 epochs, damaged, and then retrained for an additional 800 epochs. In the second, *developmental dyslexia* regime, the model was damaged prior to any training and then was trained for the full 1,800 epochs. In order to keep the computational requirements to a reasonable level, we restricted our investigation to examining the effect of damaging hidden units. We also reduced the number of sets of initial weights tested from 10 to 5, and the number of lesions at each level of lesion severity was reduced from 10 to 5. This means that each point represents the average of 25 separate trials, rather than the 100 used in previous simulations. Multiple levels of damage were investigated, ranging from 53% to 98% in steps of 3%. Figure 10 shows the results for the acquired dyslexia simulation, and Figure 11 for the developmental one. Apart from very slight differences in the number of regularization errors at low damage levels, these graphs are almost completely identical. Thus, it would seem that once the effect of relearning after damage is considered, there is no difference between the developmental and the adult cases.

DISCUSSION

We have presented four simulations using a network architecture that is essentially the same as that used in PMSP96. In the first simulation, we illustrated that under a traditional account of surface dyslexia (respecting the subtractivity assumption), the model cannot account for the pattern of deficits found in patients. In particular, the frequency × consistency interaction was not of sufficient magnitude, and the rate of regularization errors was far too low.

In the second simulation, we introduced the concept of retraining after damage, to allow the network's natural plasticity the chance to compensate for the damage—a period corresponding to the spontaneous recovery observed following brain damage in humans. With the addition of this recovery period, the behavior of the model was very similar to that found in surface dyslexic patients. The frequency × consistency interaction was found to be much larger, as was the rate of regularization errors.

In the third simulation, we modeled progressive aphasia by administering repeated microlesions, interspersed with small periods of recovery. In general, the results of this simulation were broadly in line with those from Simulation 2. However, there were two distinctive features that might help to shed light on the observation that most of the *pure* cases of surface dyslexia come from patients with a progressive disease, rather than from stroke or head injury patients. First, the magnitude of the frequency × consistency interaction was even greater for this kind of incremental damage. Second, and perhaps more important, the large magnitude of the frequency × consistency interaction was maintained over a considerable fraction of the degeneration period, whereas for damage arising from a single lesion, only a relatively small range of damage severities result in really large frequency × consistency interactions. This model thus predicts that all patients with surface dyslexia arising from progressive impairments will, at some point in the course of their disease, display very large frequency × consistency interactions. On the other hand, only a small fraction of patients with surface dyslexia arising from a single acute lesion would be expected to exhibit these very large frequency × consistency interactions, since
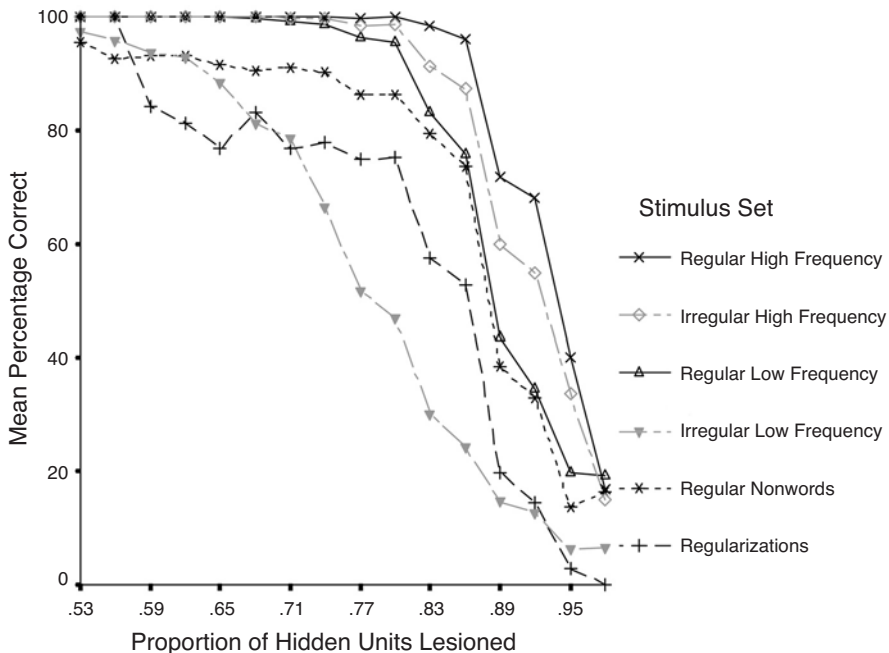

Figure 11. Developmental surface dyslexia curve, which shows the performance of the network after the sequence of damage followed by 1,800 epochs of training.

these patients would be distributed over the whole range of possible damage severities. This result is probably best explained by considering how the rate of progressive decline affects the network's ability to reach equilibrium. In the early stages of the disease, a relatively large number of links are being destroyed in each iteration of damage (approximately 100). We know from Simulation 2 that the equilibrium performance for this damage would be 100% on all word types; however, we also know that if the network is not given sufficient time to reach equilibrium, the low-frequency irregular words will be affected the worst (see Figure 6A). Toward the end of the disease, however, most of the links have already been destroyed, so that each iteration is now removing far fewer links. This means that the network is more able to reach equilibrium, which explains why performance of the other word types can be maintained at quite high levels (see Figure 6B). Thus, the key to the progressive simulation is an exponential decay rate that is initially too quick for the network to remain at its equilibrium point but that, as damage progresses, allows the network to get nearer its theoretical maximal performance.

Our final simulation explored the distinction between developmental and adult disorders. This key issue has been debated considerably, with some arguing that developmental and adult disorders are behaviorally indistinguishable, whereas others have argued that the process of developmental change must modify the nature of the deficit in the developmental case. We have demonstrated that the performance of a network damaged "at birth" and trained for 1,800 epochs was indistinguishable from the performance of a network trained for 1,000 epochs, damaged, and than retrained for a further 800 epochs. The implication of this finding is that it is legitimate to view developmental and adult disorders within the same theoretical framework—although not if that framework depends on the traditional assumption of subtractivity. This finding may have the potential to reconcile the two views described above, since it accounts for the similarity of behavioral performance in the two cases, not by ignoring the effect of development in the developmental case, but by introducing a period of "developmental" change in the adult case corresponding to the period of spontaneous recovery. It should be noted, perhaps, that this has been tested only in the domain of reading, where there is little or no age-related effect on plasticity, because of the systematic and compositional nature of the representations (Zevin & Seidenberg, 2002). It is possible, however, that developmental and adult models may diverge more in domains characterized by arbitrary mappings between input and output representations (e.g., mapping between semantics and phonology).

Another aim of this study was to explore how the concept of retraining following damage might impinge upon the traditional neuropsychological assumptions of subtractivity and transparency. We have shown that retraining after damage results in a much greater behavioral dissociation than is found immediately after damage. This implies that the reorganization of the damaged system under the influence of normal environmental learning pressures can significantly affect subsequent behavioral performance. This is difficult to reconcile with the principle of subtractivity, which assumes that the behavioral performance of brain-damaged patients can be accounted for by the operation of normal brain systems in the absence of the affected system. It seems that as the relearning process improves the network's overall performance, it also acts to inflate small postlesion processing biases into large postrecovery deficits. This occurs as the system adjusts to the demands of processing with inadequate (postinsult) resources. Under these conditions, the system will be very sensitive to variations in task difficulty, with the most difficult tasks being the least likely to recover in the face of competition for the newly impoverished resources.

Of course, in all of these simulations, we are making the assumption that there is some postinsult relearning and that this learning is similar to the original (at least in terms of the relative frequencies of the words encountered). It may well be that the rate of learning postinsult is slower or that it varies between individuals, but these considerations would not substantially affect our results. Only if there is no postinsult relearning or if a substantially different training corpus is used in relearning will our models be threatened.

These findings have two important consequences for the way in which we should interpret neuropsychological data. First, they indicate that we should be very cautious of inferences about the modular structure of cognitive processes that are based solely on evidence from performance dissociations. Even a very strong dissociation in a patient's performance may reflect only a relatively minor processing bias in the immediately postmorbid brain. This is clearly an important conclusion, since performance dissociations and, in particular, double dissociations are often regarded as the arbiter of modular structure (Ellis & Young, 1988). Our study adds further weight to the case that even very strong double dissociations cannot necessarily be regarded as evidence for two distinct and mutually exclusive processing mechanisms (Farah, 1994; Plaut, 1995). Second, these results indicate that the traditional static view of adult brain systems may need to be replaced with a more dynamic picture in which performance is viewed as the product of an equilibrium reached between three factors: computational resources, environmental learning pressures, and natural decay/noise generating processes within the brain. On this view, the process of spontaneous recovery that occurs after damage reflects the adaptation of the system as it seeks to find a new equilibrium following a significant reduction in computational resources.

Although not the key focus of this study, we should also give some thought to the implications of these results for theories of surface dyslexia. The *traditional* dual-route model of surface dyslexia assumes that surface dyslexia arises from the selective impairment of a lexical route. However, the dual-route model, even in its latest computational incarnation (Coltheart, Rastle,

Perry, Ziegler, & Langdon, 2001), does not incorporate any learning process, so it cannot account for either developmental or postdamage learning phenomena. Although the connectionist approach to models of reading is much more suited to investigating these issues, even these models have tended to adopt the traditional assumption that that performance immediately postdamage equates to patient performance in the chronic phase, and it is this assumption that leads inevitably to the incorporation of a semantic route to these models.

Following the failure of the SM89 model to account adequately for the pattern of deficits found in surface dyslexia and the success of PMSP96, it has widely been accepted that a semantic component is required to model what is occurring in surface dyslexia. Patterson, Plaut, Seidenberg, Behrmann, and Hodges (1996) wrote that "we therefore acknowledge that another 'pathway' or, as we have characterized it, another source of input to phonology appears to be necessary to model surface alexia."

In general, we would like to agree with this analysis, especially in view of the strong association between surface dyslexia and semantic impairment/anomia. However, our results do seem to indicate that surface dyslexic symptoms can arise from another scenario that does not require two pathways—namely, that of an isolated, impaired phonological pathway that has been allowed to retrain. Clearly, the significant impact of recovery on performance, as demonstrated in this relatively simple model, indicates that the concept of postdamage retraining needs to be explored within a more complete framework that encompasses phonological, orthographic, and semantic representations—a task that is well beyond the scope of the present study.

Another area in which further work would be fruitful is the study of recovery in humans. As we have indicated, there is a paucity of neuropsychological data concerning this. Yet our present work makes a number of predictions concerning the process. First, we expect that the recovery process would result in a very considerable improvement in performance; in our model, damage that initially does not reduce performance by more than 50% will recover fully, so we would expect to see performance improvements on the order of 50% in all but the most severe cases. Second, we expect that patients will tend to move from an initial state consisting of relatively undifferentiated poor word reading to one in which their performance exhibits a strong frequency $\times$ consistency interaction. Finally, we predict that the percentage of errors that are regularizations will increase very substantially over the course of the recovery period.

CONCLUSION

This study indicates that considerations of neural plasticity may have a very significant role in explaining the patterns of deficit found following brain damage or in progressive brain diseases. This is an issue that has been sidestepped for too long and now needs to be integrated into our cognitive neuropsychological framework. Here, we have only begun the process, but it is clear that any framework of this nature, which views the adult brain as a dynamical system, may well require a rethinking of some of the long-held assumptions of the discipline. In addition to challenging and revising some traditional assumptions, this line of work opens new possibilities for exploring issues of recovery and rehabilitation in patients and may point the way to a computational platform through which the effectiveness of different rehabilitation therapies may be explored.

REFERENCES

BARON-COHEN, S. (1998). Modularity in developmental cognitive neuropsychology: Evidence from autism and Gilles de la Tourette syndrome. In J. A. Burack, R. M. Hodapp, & E. Zigler (Eds.), *Handbook of mental retardation and development* (pp. 334-348). Cambridge: Cambridge University Press.

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**, 432-446.

BISHOP, D. V. M. (1997). Cognitive neuropsychology and developmental language disorders: Uncomfortable bedfellows. *Quarterly Journal of Experimental Psychology*, **50A**, 899-923.

BROCA, P. (1861). Remarques sur le siège de la faculté du langage articulé. *Bulletins de la Société Anatomique*, **6**, 330-357. Also published in H. Hécaen & J. Dubois, (Eds.) (1969), *La naissance de la neuropsychologie du langage 1825–1865* (pp. 61-91)*. Paris: Flammarion.

BUB, D., CANCELLIERE, A., & KERTESZ, A. (1985). Whole-word and analytic translation of spelling-to-sound in a nonsemantic reader. In K. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia* (pp. 15-34). Hillsdale, NJ: Erlbaum.

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

CARAMAZZA, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain & Language*, **21**, 9-20.

COHEN, J. D., DUNBAR, K., & McCLELLAND, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, **97**, 332-361.

COLTHEART, M., RASTLE, K., PERRY, C., ZIEGLER, J., & LANGDON, R. (2001). DRC: A dual cascaded model of visual word recognition and reading aloud. *Psychological Review*, **108**, 204-256.

ELLIS, A. W., & LAMBON RALPH, M. A. (2000). Connectionist networks can provide a natural account of age of acquisition effects in adult lexical processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1103-1123.

ELLIS, A. W., & YOUNG, A. W. (1988). *Human cognitive neuropsychology.* Hove, U.K.: Erlbaum.

FARAH, M. J. (1994). Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. *Behavioral & Brain Sciences*, **17**, 43-104.

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

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

JACOBS, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, **1**, 295-307.

KARMILOFF-SMITH, A. (1997). Crucial differences between developmental cognitive neuroscience and adult neuropsychology. *Developmental Neuropsychology*, **13**, 513-524.

KERTESZ, A., & McCABE, P. (1977). Recovery patterns and prognosis in aphasia. *Brain*, **100**, 1-18.

KIRAN, S., & THOMPSON, C. K. (2003). The role of semantic complexity in treatment of naming deficits: Training semantic categories in fluent aphasia by controlling exemplar typicality. *Journal of Speech, Language, & Hearing Research*, **46**, 608-622.

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

LENDREM, W., & LINCON, N. B. (1985). Spontaneous recovery of language in patients with aphasia between 4 and 34 weeks after stroke. *Journal of Neurology, Neurosurgery, & Psychiatry*, **48**, 743-748.

McCARTHY, R., & WARRINGTON, E. K. (1986). Phonological reading: Phenomena and paradoxes. *Cortex*, **22**, 359-380.

PATTERSON, K., & HODGES, J. R. (1992). Deterioration of word meaning: Implications for reading. *Neuropsychologia*, **30**, 1024-1040.

PATTERSON, K., PLAUT, D. C., SEIDENBERG, M. S., BEHRMANN, M., & HODGES, J. R. (1996). Connections and disconnections: A connectionist account of surface dyslexia. In J. Reggia, R. Berndt, & E. Ruppin (Eds.), *Neural modeling of cognitive and brain disorders* (pp. 177-199). New York: World Scientific.

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. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical & Experimental Neuropsychology*, **17**, 291-321.

PLAUT, D. C. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain & Language*, **52**, 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**, 56-115.

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

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

SEIDENBERG, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition*, **19**, 1-10.

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

SEIDENBERG, M. S., WATERS, G. S., BARNES, M. A., & TANENHAUS, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning & Verbal Behaviour*, **23**, 383-404.

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

TARABAN, R., & McCLELLAND, J. L. (1987). Conspiracy effects in word recognition. *Journal of Memory & Language*, **26**, 608-631.

TEMPLE, C. (1997). *Developmental cognitive neuropsychology*. Hove, U.K.: Psychology Press.

THOMAS, M., & KARMILOFF-SMITH, A. (2002). Are developmental disorders like cases of adult brain damage? Implications from connectionist modeling. *Behavioral & Brain Sciences*, **25**, 727-750.

WADE, D. T., HEWER, R. L., DAVID, R. M., & ENDERBY, P. M. (1986). Aphasia after stroke: Natural history and associated deficits. *Journal of Neurology, Neurosurgery, & Psychiatry*, **49**, 11-16.

WERNICKE, C. (1874). *Die aphasische Symptomen-Complex*. Breslau: Cohn & Weigart. [Trans in G. H. Eggert (Ed.) (1977). *Wernicke's works on aphasia*. The Hague: Mouton.]

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

NOTE

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