■シンポジウム 読み書きへの学際的アプローチ

Alexia and Neural Nets

K. E. Patterson*

ABSTRACT : A neural net model (developed by Seidenberg & McClelland, 1989) computes phonological representations for alphabetic letter strings. Learning, over a series of training experiences with English words, corresponds to changes in weights on connections between units at different layers in the model. After training, the performance of the model simulates many features of the word-naming performance of adult English readers. "Lesioning" the trained model also yields at least some features of one prominent form of acquired alexia. The distributed representations used by this and other neural net models have important implications for modelling cognitive impairments that result from brain lesions.

Jpn. J. Neuropsychol., 6 ; 90～99

Key Words : neural nets, alexia, reading, rehabilitation

Traditional theorizing in the field of cognitive psychology and neuropsychology relies on descriptive process models consisting of boxes and arrows. The diagram in Figure 1, for example, postulates that a written word (in the English alphabetic writing system) can be pronounced by using one of three different routines. One routine proceeds via orthographic word recognition and semantics, a second via orthographic recognition followed by word level transcoding from orthography to phonology, and a third via sub-word level transcoding. Such a model is utilized in neuropsychology by inferring, on the basis of a patient's performance, that the neural substrate corresponding to one or more boxes and/or arrows has been damaged, making a particular routine unavailable and therefore forcing the patient to rely on an alternative routine to perform a cognitive task. In the domain of acquired disorders of reading, deep dyslexia (Coltheart, Patterson & Marshall, 1980) and surface dyslexia (Patterson, Marshall & Coltheart, 1985) might be described in the framework of Figure 1 as follows : for a deep dyslexic patient, pronunciation of written words can only be accomplished by the routine involving orthographic recognition and semantics ; for a surface dyslexic patient, this same task is largely restricted to the routine involving sub-word level transcoding from orthography to phonology.

This kind of descriptive model has served a useful function in helping researchers to think clearly about the component processes required for complex skills like reading and writing. Such models have been particularly valued in neuropsychology because they represent hypotheses about which processes are truly separable, in the sense that a

1990年3月20日受理

*Medical Research Council, Applied Psychology Unit, 15 Chaucer Road, Cambridge CB 2 2EF, UK.

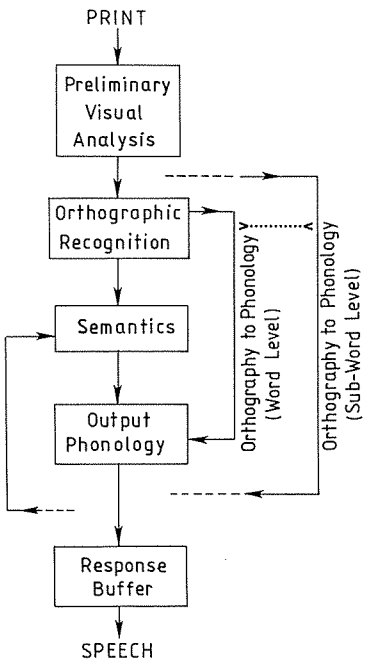

Figure 1 A descriptive model of hypothesized processes for pronouncing a written word (from Patterson, Marshall & Coltheart, 1985, General Introducition, p. XXI).

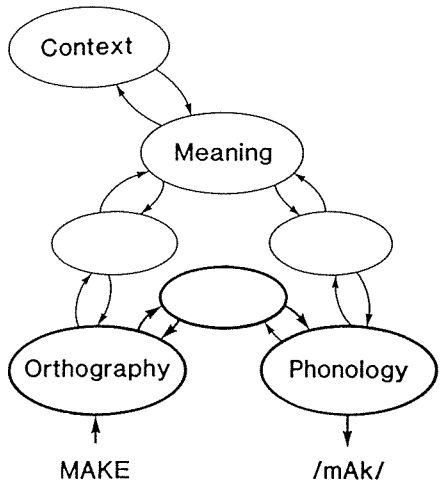

Figure 2 A general framework for orthographic, phonological and semantic processing of words ; the part of the framework in bold outline has been implemented as a neural net model by Seidenberg & McClelland (1989).

brain lesion can damage one process but leave another intact (Shallice, 1988).

Despite the recent popularity of this approach, many researchers now consider that it is limited, primarily because such a model merely states that a particular function is handled by some box or arrow, without specifying precisely how the necessary computation occurs. Neural net models are, in part, a response to this limitation. They represent one way of specifying the operation of boxes and arrows because they are working computational simulations. A neural net model does not just describe a cognitive process like the transcoding of a written word to a phonological representation : the model performs the process.

A neural net model of pronouncing written words in English has been developed by Seidenberg & McClelland (1989), and some initial explorations of this model with regard to acquired disorders of reading have been performed by Patterson, Seidenberg & McClelland (1989). The general framework of the model is shown in Figure 2, where it can be seen that, unlike Figure 1 with its three routines for written word naming, this model postulates only two : one direct computation from orthography to phonology, and one indirect computation to phonology via meaning or semantics. Furthermore, although the theory acknowledges the necessity for both of these routines, the initial implementation of the model includes only the processes in bold outline — that is, only the direct computation of phonology from orthography.

Like most neural net or connectionist models, this one consists of simple processing elements (units) with connections between them. As Figure 2 indicates, this network has three levels or layers of units :

in the simulations to be described here, 400 units code the orthographic information in a word presented to the model, 460 units code the phonology of the pronunciation computed by the network, and 200 units intervene at a "hidden" layer. These are called hidden units because they do not directly reflect any features of the real world (such as orthography or phonology) but are completely internal to the functioning of the model. In this neural net, units within a single layer are unconnected and therefore do not influence one another's operation, but there is complete connectivity between layers : every unit at the orthographic input level is connected to every hidden unit, and every hidden unit is connected to every unit at the phonological output level.

Details about the orthographic and phonological representations can be found in Seidenberg & McClelland (1989). For present purposes, it is sufficient to note just a few facts. The order of both letters and phonemes is represented not in terms of explicit position, but rather in terms of relative position : the letter K in the written word MAKE is coded not as the letter K in position 3 but as the letter K preceded by A and followed by E ; likewise /k/ in the spoken word "make" is coded as the phoneme /k/ preceded by the sound of a long /A/ and followed by nothing (a word boundary). The other thing to be noted here is that phonological units in the model do not actually correspond to triplets of phonemes, but rather to triplets of phonetic features of phonemes such as place of articulation, voicing, etc. This phonological coding scheme was borrowed from the past-tense verb learning connectionist model developed by Rumelhart & McClelland (1986).

The current network was trained on a vocabulary of almost 3000 words, which is not all but most of the monosyllabic and monomorphemic words in the English language. Training was carried out in a series of "epochs" ; during each epoch, around 500 words were presented to the net for processing. The items were selected from the vocabulary at random but modulated as a function of word frequency, with the function closely related to the log of word frequency as tabulated by Kucera & Francis (1967).

Connections between units in the different layers carry weights which are given random values at the beginning of training. Over the series of training trials, the weights on these connections are gradually altered, using the back-propagation learning algorithm (Rumelhart, Hinton & Williams, 1986), to reduce the discrepancy between the model's computed pronunciation for a letter string and the correct pronunciation.

In fact, this simulation model of written word pronunciation does not pronounce words : it has not been implemented with a voice. The model simply computes a pattern of activation across the 460 phonological units, where each unit is either "off" or is activated to some degree approaching maximum. The performance of the model can be evaluated in one of two ways. The qualitative measure attempts to determine whether the pattern of activation across the phonological units computed by the model is closer to the correct pattern for the presented word than to the pattern for any other word or string. This measure is similar to assessing a human subject's accuracy (per cent correct) in naming printed words. It would be cumbersome to compare the model's output to a very large number of

alternative patterns ; the current version of the simulation (see Seidenberg & McClelland, 1989 for details) compares the model's output to the correct pronunciation and to any pronunciation deviating from the correct one by a single phoneme, and reports the best fit. The second performance measure is quantitative : for each phonological unit, the discrepancy in activation level between expected (what the model would compute if it were performing perfectly) and observed (what the model has actually computed) is squared and the squared difference values from all phonological units are then combined to give an "error" score. The advantage of this quantitative performance measure is that two words which are both correct (in the qualitative sense noted above) can still yield different error scores. A low error score can be thought of as the model's equivalent of quick, efficient, noise-free processing. Error scores are intended to provide a simulation of real subjects' reaction times in naming written words.

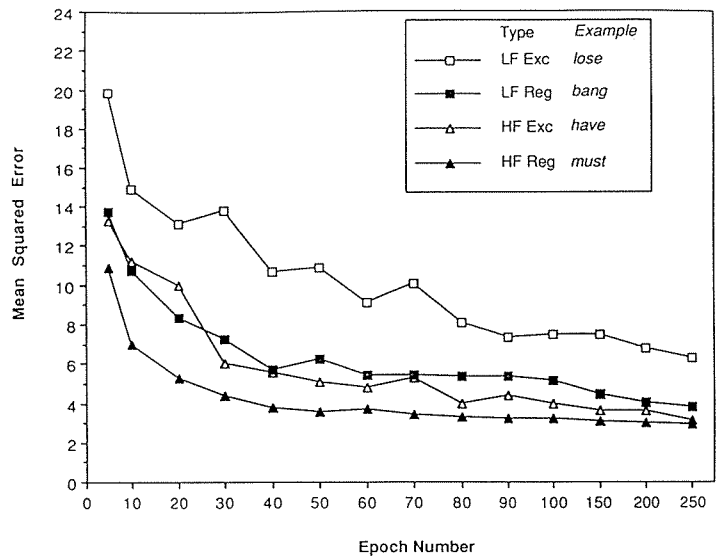

Figure 3 Performance of the neural net model on four types of words over a series of training epochs (from Seidenberg & McClelland, 1989, p. 535).

Figure 3 shows the performance of the model over 250 training epochs on a set of words from an experiment by Taraban & McClelland (1987). The stimulus list contained 96 words, 24 in each of four classes : high frequency regular words (like the word MUST) are English words that occur very commonly and also have a regular spelling-to-sound relationship ; high frequency exception words (e.g. HAVE) are also common individual items but do not exemplify the common pronunciation of their spelling patterns (SAVE, GAVE and WAVE are regular examples contrasting with HAVE); low frequency regular words (like BANG) are less commonly encountered items with regular spelling-to-sound correspondences; and low frequency exception words (LOSE) have both lower familiarity as whole items and also embody an atypical correspondence (regular examples of this spelling pattern are POSE and HOSE).

Figure 3 reveals that performance as measured by the quantitative error score improves over the training period for all four word classes. How do the independent variables of word frequency and spelling-to-sound regularity influence the model's performance at various stages of training? Early in training (for example, at epoch 20), there are major effects of both frequency and regularity : the low frequency words, represented by squares, have larger error scores than the high frequency words,

represented by triangles ; and the exception words, represented by open symbols, have larger error scores than the regular words, represented by filled symbols. This pattern is of interest because it corresponds to the characteristics of word naming performance by children who are learning to read in English : as demonstrated by Backman, Bruck, Hebert & Seidenberg (1984), children's success in reading aloud is facilitated by both familiarity and regularity. Late in training (epoch 200, for example), there are no longer two independent effects but rather an interaction : for low frequency words (the squares), regularity still influences the model's performance ; but for words on which the model has had many training experiences (high frequency words, the triangles), there is no longer a significant advantage for items with regular correspondences. This interaction is of interest because it is characteristic of word naming performance by adult skilled readers of English (Seidenberg, Waters, Barnes & Tanenhaus, 1984 ; Taraban & McClelland, 1987). The model shows a close, quantitative simulation of the effects of these two stimulus variables on word naming by both novice and experienced readers of English. Simulations of other effects can be found in Seidenberg & McClelland (1989).

As indicated earlier in the paper, the broader theory of which the implemented model is a part postulates two routines by which a reader can compute the pronunciation of a written word. In attempting to capture a whole range of human performance characteristics with the implementation of just one of these routines, the model represents a bold hypothesis that the majority of word naming is handled by this single procedure. Although more traditional information processing models like the one displayed in Figure 1 have been somewhat agnostic about the role of the semantic routine in normal word naming, the proposers of such models have thought it necessary to postulate two other, more-or-less independent, procedures : one involving "addressed" (word level) and another involving "assembled" (sub-word level) phonology (see Coltheart, 1980, 1987 for reviews of the evidence underlying this distinction). One of the major sources of this evidence has been studies of neurological patients with acquired disorders of reading.

Of particular relevance to the present neural net model is the form of acquired dyslexia known as "surface" dyslexia (Marshall & Newcombe, 1973). In its purest form (see Bub, Cancelliere & Kertesz, 1985 ; McCarthy & Warrington, 1986 ; Shallice, Warrington & McCarthy, 1983), this reading disorder is characterised by (a) relatively fluent and correct naming of words with a regular spelling-to-sound correspondence (e.g. MUST and BANG from the list in Figure 3) ; (b) a significantly higher error rate on words with an exceptional spelling-to-sound correspondence (e.g. HAVE and LOSE) ; and (c) a predominance of one specific error type on exception words : the patient's response represents a "regularized" or typical pronunciation of the spelling pattern (e.g. LOSE pronounced to rhyme with "pose" or "hose").

The general question to be addressed now is whether the model, so successful in accounting for word naming by normal readers, can also advance our understanding of the breakdown of this skill as a consequence of neurological damage. The specific question is whether some form of "lesion" or damage to the trained network

can simulate the characteristics of surface dyslexia mentioned above. There are two major reasons for selecting the "pure" form of surface dyslexia as a test of the model's ability to capture impaired as well as intact word naming performance. First of all, such patients have severely impaired word comprehension, suggesting that their word naming performance is largely uninfluenced by the routine, available to normal readers, in which the computation of phonology from orthography is mediated by word meaning. Since the neural net also knows nothing of word meanings, its parallel to surface dyslexia has at least face validity.

The second point concerns the striking claim of this neural net model that a single procedure for the direct computation of phonology from orthography can handle all types of words and letter strings in English. As already noted, traditional information processing models have not managed without two different procedures to explain a number of features of both normal and impaired performance, and surface dyslexia has been one of the sources of evidence claimed to support this multiple-routine view. It is simple enough to "explain" the pattern of reading performance in surface dyslexia in terms of a model like Figure 1 : an English reader forced (by damage to alternative procedures) to rely primarily on an intact sub-word level routine should name regular words correctly but give regularized pronunciations to exception words. A challenge for a single-process account like the neural net model is whether some form of disruption to the single computational process can also produce the surface dyslexic pattern of reading.

The trained network can be disrupted in various ways : processing structures can be eliminated or noise can be added to the processing procedures. It can also be damaged in various locations : in the present model, for example, either hidden units or connections from hidden units to phonological output units could be disrupted. Finally, the net can be lesioned to varying degrees: different proportions of units or connections can be affected. Some of the features associated with these options for damaging the network have been explored in Patterson, Seidenberg & McClelland (1989). Here, just one lesion experiment will be reported, in which 20% of hidden units were damaged: that is, after training to 250 epochs, the activation values of 40 hidden units (selected at random from the 200 units at the hidden layer) were fixed at zero so that they could not contribute to the model's computations. The model was tested on the Taraban & McClelland (1987) high and low frequency regular and exception words 10 times, each time with a new random 20% of the hidden units zeroed. The model's damaged performance, averaged over the 10 tests, was then compared to the performance of two surface dyslexic patients asked to name comparable words. The two patients selected from the literature were KT, a patient with a diagnosis of pre-senile dementia (McCarthy & Warrington, 1986) and MP, a patient with severe left-temporal damage following a road traffic accident (Bub, Cancelliere & Kertesz, 1985).

It is important to note that, since (a) the patients were studied prior to the development of the model, and (b) the precise word-sets used with patients are not always reported in detail, it was not possible to compare network and patients on identical lists. One of the impressive features of Seidenberg & McClelland's evaluation

Table 1 Word naming performance (per cent correct) on high and low frequency regular and exception words by two surface dyslexic patients (KT : McCarthy & Warrington, 1986 ; MP : Bub, Cancelliere & Kertesz, 1985) and by the neural net model with 20% of its hidden units "lesioned".

	High Frequency		Low Frequency	
	Regular	Exception	Regular	Exception
KT	100	47	89	26
MP	95	93	98	73
Model	93	86	93	78

of their computational model with respect to normal readers is the use of identical stimulus materials. Assessment of future surface dyslexic patients will permit tighter control in this regard ; for the moment, it is merely possible to assess performance on similar classes of words. Data for MP were taken from Bub et al (1985, Table 1. 2, p. 21) ; performance for KT (R. McCarthy, personal communication) includes virtually all of the Taraban & McClelland exception words on which the model was tested, but somewhat different (though comparable) lists of regular words.

Table 1 shows the performance of the patients. As always in neuropsychology, there is considerable variation amongst patients with the same general symptom complex : KT was dramatically poorer at naming exception than regular words, even for high frequency words. For MP, the regularity effect was restricted to lower frequency words. In fact, MP basically shows in accuracy of word naming what normal adult readers show in speed of word naming : an interaction between regularity and frequency effects.

The performance of the model with 20% of hidden units lesioned is also shown in Table 1. The model shows a small effect of ragularity for high frequency words and a somewhat larger effect for low frequency words : essentially, a rather good match to MP's performance. The lesioned model represents a less satisfactory simulation of KT's very substantial regularity effects. It might seem an obvious step (towards a closer simulation of KT) to inflict a greater degree of damage on the net, but the model's performance on regular words begins to decline when larger proportions of hidden units are eliminated. Thus far, our lesioning explorations have not reproduced the dramatic pattern of performance shown by KT.

What about the other major feature of surface dyslexic performance noted above, the predominant type of error on exception words? Since the simulation compares the model's computed pattern of activation over the phonological output units to each pattern differing from the correct one by a single phoneme and reports the best fit, one can assess whether the model's "pronunciation" errors are a good match for the patients' errors. Across a large number of exception words, both MP and KT produced the regularized pronunciation (e.g. PINT named as if it rhymed with "hint" and "mint") on about 85% of error responses. The other 15% of their erroneous responses were pronunciations with some orthographic/phonological resemblance to the target word (e.g. PINT named as "paint"). The lesioned model's errors on the Taraban & McClelland exception words were split almost exactly 50/50 between exact regularizations and these other kinds of pronunciation errors. Thus the network successfully simulates the occurrence of the right kinds of errors in surface dyslexia, though not necessarily the precise proportions of these error types that are bserved

in specific patients.

These preliminary results suggest a promising start in considering alexia within the neural net approach. There is much work to be done, both to improve the simulation of surface dyslexia and to attempt the simulation of other reading disorders such as deep and phonological dyslexia. Detailed modelling of these other varieties of alexia will require the implementation of other parts of the theory (shown in Figure 2 but not in bold outline) : the transcoding procedure from orthography to meaning (see Hinton & Shallice, 1989, for a neural net model of this computation) and thence from meaning to phonology.

Since this paper originally formed part of a symposium on various approaches to the study of alexia, it seems appropriate to conclude with a comment on an aspect of the neural net approach which has implications for our general understanding of cognitive disorders and their rehabilitation. Like many neural net models, this one embodies the principle of distributed representations. As contrasted with local representations, where one meaningful entity (for example, the phonological form of a particular word) is represented by one element or location in the model, here a word is represented by many different units, and many different connections are involved in its processing. Likewise, no given unit or connection in the model can be considered to belong to a particular word: rather that element or connection participates in the representation of and processing of many different words.

The consequences of distributed representations for lesioned performance seem to fit the behaviour of real patients in several important ways. One of these is so-called "graceful degradation" with damage : as the model is incrementally damaged, its performance degrades gradually rather than in an all-or-none fashion. Most neuropsychologists would concur with Allport's (1985) observation that the performance of brain-lesioned patients on cognitive tasks is also not all-or-none : the most typical feature of performance by an aphasic or alexic patient trying to compute the pronunciation of a word is that performance is less efficient and less reliable than normal. The patient is likely to be slow, and may achieve only a partially correct representation, or may indeed fail altogether ; but on another occasion, the patient may succeed on that same word. As Allport (1985) emphasises this pattern fits neatly with the idea of distributed representations: since no one element or connection is essential to success on a particular item, partial damage to the network should result in precisely this observed variability.

The second point involves generalization of learning, whether this is initial learning or re-learning in treatment for cognitive impairment. In the neural net, all learning corresponds to changes in the weights on connections. Since any given connection is involved in processing many different words, it is a prediction of this kind of model that there should always be some degree of generalization of learning or re-learning (Hinton, McClelland & Rumelhart, 1986). Training on one particular group of items (say, set A) should have maximum implications for the processing of items in group A, because it is the connections specifically relevant to A items which are being exercised and altered ; but since those same connections are also used to some extent by other, similar items (in set B), the effects

of training on A items should generalise to some degree to untreated B items. Coltheart & Byng (1989) have shown precisely this pattern of generalization effects in their treatment of a surface dyslexic patient (see also Wilson & Patterson, 1990, for further discussion).

Generalization of treatment effects indicates the relevance of neural net models to rehabilitation only at a very general level. One can, however, hope for a future stage offering more detailed interaction between theory and therapy. As neural net models grow in scope and detail, neuropsychologists may be able to use these models as a source of hypotheses as to which items and which types of training might provide maximum generalization and benefit.

Acknowledgements : This paper is based on a contribution to a symposium at the 13th Annual Meeting of the Neuropsychology Association of Japan, Tokyo, September 1989. I am grateful to the President of the meeting (Dr. S. Sasanuma) and to the Chairman of the symposium (Dr. M. Iwata) for giving me the opportunity to participate. I thank Dr. R. McCarthy for access to data from patient KT.

REFERENCES

1) Allport, D. A. : Distributed memory, modular subsytems and dysphasia. In Current Perspectives in Dysphasia (ed. by Newman, S. K. and Epstein, R.), Churchill Livingstone, Edinburgh, 1985.

2) Backman, J., Bruck, M., Hebert, M. and Seidenberg, M. S. : Acquisition and use of spelling-sound information in reading. Journal of Experimental Child Psychology, 38 ; 114-133, 1984.

3) Bub, D., Cancelliere, A., and Kertesz, A. : Whole-word and analytic translation of spelling to sound in a non-semantic reader. In Surface Dyslexia (ed. by Patterson, K., Marshall, J. C. and Coltheart, M.), Erlbaum, London, 1985.

4) Coltheart, M. : Reading, phonological recoding, and deep dyslexia. In Deep Dyslexia (ed. by Coltheart, M., Patterson, K. and Marshall, J. C.), Routledge & Kegan Paul, London, 1980.

5) Coltheart, M. : Cognitive neuropsychology and the study of reading. In Attention and Performance XI (ed. by Posner, M. I. and Marin, O. S. M.), Erlbaum, Hillsdale NJ, 1985.

6) Coltheart, M. and Byng, S. : A treatment for surface dyslexia. In Cognitive Approaches in Neuropsychological Rehabilitation (ed. by Seron, X. and Deloche, G.), Erlbaum, Hillsdale NJ, 1989.

7) Coltheart, M., Patterson, K. and Marshall. J. C. : Deep Dyslexia. Routledge & Kegan Paul, London, 1980.

8) Hinton, G. E., McClelland, J. L. and Rumelhart, D. E. : Distributed representations. In Parallel Distributed Processing. Volume 1 (ed. by Rumelhart, D. E. and McClelland, J. L.), MIT Press, Cambridge Mass, 1986.

9) Hinton, G. E. and Shallice, T. : Lesioning a connectionist network : investigations of acquired dyslexia. University of Toronto (Department of Computer Science), technical report CRG-TR-893.

10) Kucera, H. and Francis, W. N. : Computational Analysis of Present-Day American English, Brown University Press, Providence RI, 1967.

11) McCarthy, R. and Warrington, E. K. : Phonological reading : phenomena and paradoxes. Cortex, 22 ; 359-380, 1986.

12) Marshall, J. C. and Newcombe, F. : Patterns of paralexia : a psycholinguistic approach. Journal of Psycholinguistic Research, 2 ; 175-199, 1973.

13) Patterson, K, Marshall, J. C. and Coltheart, M. : Surface Dyslexia. Erlbaum, London, 1985.

14) Patterson, K, Seidenberg, M. S. and McClelland, J. L. : Connections and disconnections : acquired dyslexia in a computational model of reading processes. In Parallel Distributed

Processing : Implications for Psychology and Neurobiology (ed. by Morris, R. G. M.), Oxford University Press, Oxford, 1989.
15) Rumelhart, D. E., Hinton, G. E. and Williams, R. J. : Learning internal representations by error propagation. In Parallel Distributed Processing, Volume 1 (ed. by Rumelhart, D. E. and McClelland, J. L.), MIT Press, Cambridge Mass, 1986.
16) Rumelhart, D. E. and MuClelland, J. L. : On learning the past tenses of English verbs. In Parallel Distributed Processing, Volume 2 (ed. by McClelland, J. L. and Rumelhart, D. E.), MIT Press, Cambridge Mass, 1986.
17) Seidenberg, M. S. and McClelland, J. L. : A distributed, developmental model of word recognition and naming. Psychological Review, 96 ; 523-568, 1989.
18) Seidenberg, M. S., Waters, G. S., Barnes, M. A. and Tanenhaus, M. K. : When does irregular spelling or pronunciation influence word recognition? Journal of Verbal Learning and Verbal Behavior, 23 ; 383-404, 1984.
19) Shallice, T. : From Neuropsychology to Mental Structure. Cambridge University Press, Cambridge, 1988.
20) Shallice, T., Warrington, E. K. and McCarthy, R. : Reading without semantics. Quarterly Journal of Experimental Psychology, 35A ; 111-138, 1983.
21) Taraban, R. and McClelland, J. L. : Conspiracy effects in word recognition. Journal of Memory and Language, 26 ; 608-631, 1987.
22) Wilson, B. and Patterson, K. : Rehabilitation for cognitive impairment : does cognitive psychology apply? Applied Cognitive Psychology, in press, 1990.