EFFECTS OF DAMAGE TO THE CDM STROOP MODEL

J. WILES*[†], H. CHENERY[‡], J. HALLINAN[*], A. BLAIR[A] AND D. NAUMANN[B]

**School of Computer Science and Electrical Engineering, University of Queensland, Brisbane*
[†]School of Psychology, University of Queensland, Brisbane
[‡]Department of Speech Pathology and Audiology, University of Queensland, Brisbane
[a]Department of Computer Science, University of Melbourne
[b]Boeing Australia, Brisbane

We report simulations of Cohen, Dunbar and McClelland's (CDM) model of the Stroop effect showing how damage to the network can replicate empirical data on older individuals and those with dementia of the Alzheimer's type (DAT). The study is significant on three counts: firstly, the specific patterns of damage required are theoretically justifiable, leading to a deeper understanding of the role of connectionist models in understanding DAT; secondly, the simulations show that in specific neurodegenerative conditions, the CDM model can replicate behavioral effects, providing increased confidence in the model and countering recently published claims to the contrary; and thirdly, a new component of the model, the preliminary time constant, is identified as a major factor in the ability of our damaged model to successfully replicate empirical data, indicating a potential role for deficits in attention to task instructions.

1 Introduction

Connectionist models of human information processing and the simulation of damage to these models are powerful tools for investigating and understanding specific patterns of cognitive impairment associated with neurodegenerative diseases such as Alzheimer's Disease. Laboratory based experimental tasks (such as the Stroop task) have provided important clinical measures of patient behaviour. The computational models provide a complementary methodology to investigate assumptions made explicit by theories of normal and impaired cognitive processing. In this project we present a study of Cohen, Dunbar and McClelland's (CDM) neural network model of the Stroop effect [1] subjected to different damage techniques. We compare the resulting behaviour with results from the literature on the performance of patients with Dementia of the Alzheimer's Type (DAT).

1.1 Modelling Automaticity and Control in Alzheimer's Disease

Many types of skilled activities such as language and information processing reflect both automaticity and control. Automaticity is reflected by the fact that a task draws nothing (or very little) from limited mental resources, and control is manifested by the ability to ignore a stimulus or inhibit its processing as the prevailing conditions (or situations) require. In connectionist terms, inhibition can be equated with the components of a neural net (such as negative weights and biases) that reduce or stop processing of distractor items (see [4]). The behavioural consequences of inhibition have been reported in a variety of effects on the Stroop task where the processing of an unattended stimulus is measured by the effect of that stimulus on a concurrent target (see [6] for a review of the Stroop effect).

In the Stroop task, subjects are asked to perform both a word reading and a color naming task under three conditions. In the congruent condition, words (e.g., blue) appear in their corresponding ink color. In the incongruent condition, a color name (e.g., blue) appears in a nonmatching color. The time taken to either read the word or name the color in which the word appears is compared with a neutral condition. The neutral condition for the color naming task may entail presenting color patches or rows of colored XXXX's to be named, or presenting neutral words (e.g. “empty” or “deep”) in various colors to be named. The neutral condition for the word naming task typically uses the color names presented in black ink but may also use other words. The critical measures are (a) the processing advantage for the congruent when compared with the neutral condition (termed facilitation) and (b) the interference condition (incongruent compared with neutral condition) which has been used as an indicant of the efficiency of the inhibitory system.

Several accounts of the Stroop effects have been proposed but the CDM model was a particularly important milestone as it proposed that automaticity is a matter of degree, and that tasks such as word reading and color naming lie along a continuum of automaticity that is dependent upon the relative degree of automatization of each task. In general terms, the CDM model consists of two pathways, with each pathway consisting of a set of input, intermediate (or hidden) and output units. In addition to the two pathways, there are two task demand (or attention) units - one for the color naming task and the other for the word reading task. In the network, these inputs are connected to the intermediate units in the two processing pathways and have an effect similar to the allocation of attention to one or the other of them (see Figure 1).

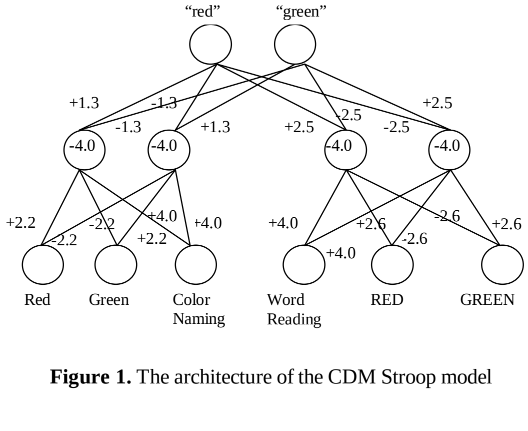

Figure 1. The architecture of the CDM Stroop model

Whilst Cohen *et al*'s model depicting the functional architecture of the Stroop task has provided an integrated theory with which to understand mechanisms of inhibition in normal subjects, what is needed for progress in the area is to extend the predictions of the model to account for neurodegenerativ e damage.

1.2 The Stroop Effect in Alzheimer's Disease

In recent years, various researchers (e.g. [3]) have proposed that at least some of the cognitive deficits manifest by patients with DAT may have their basis in disturbed facilitatory and inhibitory processing, or a combination of both. The Stroop task is a useful experimental paradigm with which to examine the nature of the neurocognitive deficits in DAT because it allows for the measurement of both facilitatory and inhibitory processes through reaction time recording.

A detailed study of the Stroop effect in DAT was undertaken by Spieler, Balota and Faust [8]. Their major findings included: (a) in the color naming task, there was a significant increase in facilitation (termed hyperfacilitation) for subjec ts with very mild DAT, and (b) there was no reliable increase in interference effects for the DAT group. Spieler *et al.* showed, however, that the increased difficulty that patients with mild DAT experience when they are presented with conflicting information is manifested in their large increase in intrusion errors (when subjects actually read the word rather than name the color on the color naming conflict condition) when compared with the control groups. These errors on the incongruent color naming task were typically made quickly. Spieler *et al.* proposed that “the results are constant with an inhibitory breakdown in normal ageing and an accelerated breakdown in inhibition in DAT individuals” (p. 461). This observation provides a baseline theory which can be used to guide the choice of damage parameters in a model of the system.

A final behavioural effect common to most brain damaged populations is that the DAT individuals had significantly slower RT's across all conditions. This observation, whilst reported frequently in the clinical literature, is important to mention in a modeling study, as one of the damage techniques (described below) showed qualitatively similar behaviours to the former findings without a major increase in the mean RT's.

2 Simulation of Damage in the Stroop Network

2.1 Aims and Approach

The aim of this study is to model Spieler *et al.*'s [8] data. Ours is not the first attempt: Kanne, Balota, Spieler and Faust [5] attempted to model the DAT data, without success, as they encountered problems in modifying the CDM architecture to incorporate set size effects. Their difficulties are discussed elsewhere [5], together with suggestions by Cohen, Usher and McClelland [2] for alternative modifications to achieve their set-size goals.

In this study, we have followed a different methodology from Kanne *et al.* [5], retaining the initial CDM architecture and adapting it to the Spieler *et al.* [8] data, rather than modifying the architecture itself. From an initial set of weights which produced behavior matched to Spieler *et al.*'s young subjects, we then studied the effects of damage to the weights of the network and modifications to the parameters of the model. The results of these damage conditions were compared to those of the DAT individuals and the older control subjects.

2.1.1 Replication: discovering the effects of τ_0

The first stage of our study involved simulating the original CDM network. The simulation parameters and procedures were as reported in Cohen *et al.* [1], except for two aspects: (a) The standard deviation of the noise in the evidence accumulators was set to 0.01 (instead of 0.1, correcting an error noted in [9, p. 874 footnote 1]) and (b) a second timing parameter, τ_0 was introduced: Cohen *et al.* [1] reported allowing the network to "settle" prior to applying the conditions of the task. This preliminary settling stage is a component of the initial instructions to a subject (i.e., to focus on the color of the ink, or to name the word). Using the reported value of $\tau = 0.1$ and allowing unlimited settling time, our initial attempts at replication produced lower interference than that reported by Cohen *et al.* [1]. We tested a range of settling times, and found that 12 settling steps best approximated Cohen *et al.*'s results[1] [1]. Since the preliminary settling time had a noticeable effect on behaviour, we introduced a separate parameter, τ_0, to allow for variations in the initial attention subjects paid to the task instructions. For the remainder of the simulations, we used 12 steps of settling with τ_0=0.1, and in the damage conditions tested a variety of values for τ_0.

2.1.2 Finding weights to match the young controls

The second stage of the study aimed to reproduce the data of Spieler *et al.*'s young control subjects. Spieler *et al.*'s experimental conditions for the neutral word naming condition differed from other Stroop studies. The neutral words they used took longer for subjects to read than even the incongruent word condition, conflicting with the empirical studies that the CDM model was designed to replicate.

In order to reproduce Spieler *et al.*'s control conditions, we used a hill-climbing technique to adapt CDM's trained weights. The selection criterion was based on a least squares match to the means and standard deviations of five of Speiler *et al.*'s conditions (omitting the anomalous word naming control condition). The first champion network was defined to be the CDM network. At each generation, a

[1] This was also the number of preliminary steps used by Mewhort *et al.* [7].

mutant network was created by adding Gaussian random noise to the weights of the current champ network (but preserving the pattern of weight sharing). The champ and mutant were then run 100 times in each of the five conditions, recording the number of cycles for the accumulator to reach threshold[1]. The means and standard deviations of these numbers were then converted to reaction times (in milliseconds) by linear regression. The technique of adapting the network directly to the young controls has the advantage of providing the closest possible match between the empirical data and the starting conditions of the ne twork.

The parameters and weights found by the hill-climbing process are interesting in several respects. Firstly, a higher value was found for the positive attentional weights (5.3 compared to 4.0), and the matched negative biases. These values have the effect of reducing the leakage between channels in the model, thus reducing both the interference and facilitation effects. Secondly, values of the input-to-hidden weights were substantially differentiated, with the color channel reduced, and the word channel increased. This effect counters that of the increase in attentional values, increasing both facilitation and interference in the color-naming condition, but not the word-reading conditions. The third effect was to reduce the differentiation between the hidden-to-output units.

The hill-climbing weights and biases produced a good fit to Spieler *et al.*'s [8] young control subject data ($R^2 = 0.99$), indicating that the model can replicate the data obtained from the young subjects.

2.1.3 Effects of damage to the modified Stroop model

We tested a variety of forms of damage in the network: general semantic damage; damage to the inhibitory component of the attentional system; general slowing and partial attention deficit to the task, implemented as slowing of the update procedure during the initial settling phase[2]. For each damage condition (4 forms of damage x 5 or more parameter variations for each condition), 50 trials were run. Reaction times within each condition were averaged and the interference and facilitation effects calculated. Each damage condition had a clear effect, as described below, but the major finding is that none of the damage conditions alone replicated the empirical data for either the older or DAT individuals in Spieler *et al.*'s studies.

1. General semantic damage: Semantic damage was implemented as a percentage decrease in the magnitude of all weights (but not biases) in the network.

Results: As expected, lower weight values throughout the network produced a general increase in reaction times. W ith increasing levels of damage, reaction times increased for all conditions, but both interference and facilitation were slightly

[1] In order to make the comparison as fair as possible, the same 500 str eams of pseudo-random numbers are used for the mutant and the champ.

[2] The time course is implemented using the update parameters τ and τ_0 with each hidden- and output unit using the update equation $a(t+1) = (1-\tau) \times a(t) + \text{netinput}$. τ is replaced by τ_0 in the initial settling phase of each trial.

reduced. Reduction in interference and facilitation would be expected from a mismatch between the positive weights from the attentional system (which were damaged in the general semantic damage) and the negative biases which were not modified in this condition.

2. Inhibitory component of the attentional system: Inhibitory damage was implemented as percentage decrease in the magnitude of the negative biases in the hidden units.

Results: With increasing damage to the biases, reaction times increased marginally in the control condition and markedly in the incongruent condition. Reaction time for the congruent condition decreased slightly. The overall effect was an exaggerated interference effect and a steady increase in facilitation with increased damage. Analysis of the network shows a clear mechanism for how the damage to the attentional component of the network affected its performance. The positive task weights and negative biases work together so that when a task unit (such as the colour naming unit) is activated, one set of hidden units (such as the colour channel) would be most sensitive to its input values (the colour input units) and the other set of hidden units (the word channel) would be strongly inhibited due to their negative biases.

3. General slowing in the course of processing was modeled as a decrease in the timing parameters τ and τ_0 together. In addition, increases (corresponding to decreasing the time course) were tested for completeness.

Results: Reaction times increased systematically across all conditions inversely with the timing values. Interference also increased systematically, but proportional to the values. Facilitation showed a U-shaped pattern, with the minimum facilitation at 0.1.

4. Less attention to the task instructions was implemented as a decrease in τ_0. In addition, increases were tested for completeness.

Results: Decreasing τ_0 had no effect on the control reaction time, but did increase both interference and facilitation. Increasing τ_0 caused no systematic variation in reaction times, interference or facilitation effects.

The individual damage studies had a variety of effects, some consistent with different aspects of Spieler *et al.*'s data, but none with both increased RTs, and proportionately increased facilitation without exaggerated interference. Since effects interact within the model, simple linear combinations did not approximate the empirical behaviour.

2.1.4 Effects of combined forms of damage.

In the final stage of the study, we tested a variety of combined forms of damage, chosen initially based on the results of the previous analyses and modified according to Spieler *et al.*'s theories of damage. The network weights were originally fitted to the young condition, and therefore the parameters originally used in the CDM model were retained unaltered. By combining different types of damage it was possible to model Spieler *et al.*'s empirical data very closely (Figure 2). The parameter settings used for these networks are shown in Table 1.

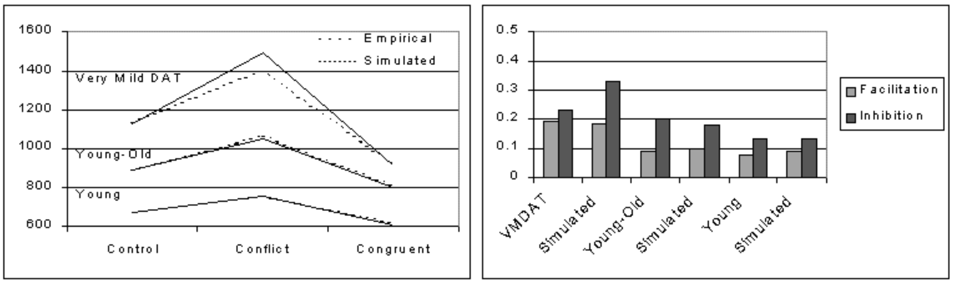

Figure 2. Combining different types of damage to model empirical data

Table 1. Parameter settings for Stroop DAT model

CONDITION	τ_0	τ	WEIGHT DAMAGE	BIAS DAMAGE
Young	0.100	0.100	0.000	0.000
Young-Old	0.500	0.500	0.040	0.040
Very Mild DAT	0.040	0.042	0.170	0.210

The young-old condition is characterized by a general slowing of response time, and a slightly increased level of interference (from 0.13 to 0.20). This condition was modelled by damaging the weights and biases by a small amount (4%), in conjunction with a reduction in τ and τ_0 from 0.1 to 0.05. That is, the pattern of damage is consistent with a slight general slowing down of the system (τ), reduced efficiency of processing (general damage) together with a reduction in the initial attention mechanism (τ_0).

The very mild DAT condition exhibits still further slowing of response times, slightly increased interference, and hyperfacilitation. In the model, this pattern of behaviour was achieved by further reducing τ, and τ_0, with τ_0 slightly lower than τ. Damage to the weights and biases was increased, with proportionally greater damage to the inhibitory biases on the hidden units than to the weights. The control and congruent response times were fitted well by the model, but the incongruent response time was consistently too high. On further analysis of the model, this effect was found to be the result of the manner in which the output accumulator operates. Evidence is added to the accumulator in the form of small random numbers drawn from a distribution with a mean equal to the difference between the activation values of the two output units of the network. This means that when one output unit is strongly activated, and the other only weakly, a fairly large positive amount will be added to one accumulator, and an equal negative amount to the other. Under these conditions it is nearly impossible for intrusion errors to occur – that is, the "wrong" accumulator cannot reach threshold just because the "right" one is taking a long time to increase. Intrusion errors can be modelled to some extent by setting a restrictive timeout on the system, so that if no accumulator has reached threshold within a set time an intrusion error is deemed to have occurred. This approach gives the right shape to the response time curves, but for the wrong reason, as pointed out by Mewhort *et al.* [7].

The pattern of damage necessary to produce very mild DAT-type behaviour suggests that such Stroop behaviour may result from a combination of general slowing, as seen in healthy Young-Old individuals, but at a higher level, together with additional damage to the inhibitory system and to the attentional system.

3 Conclusions

This study supports the utility of the CDM Stroop model, and demonstrates how it can be used to model empirical observations regarding Stroop behaviour in individuals of different pathologies. The most serious drawback apparent is the use of an output accumulator to model response times, a subsystem which does not permit the modelling of intrusion errors, which become significant in DAT sufferers.

The model supports the hypothesis that DAT involves damage to the inhibitory part of the attentional system, and further suggests that initial attention to the task has an important effect upon facilitation and interference.

References

1. 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.
2. Cohen, J. D., Usher, M. & McClelland, J. L. (1998). A PDP approach to set size effects within the Stroop task: Reply to Kanne, Balota, Spieler and Faust (1998). Psychological Review, 105, 188-194.
3. Faust, M.E., Balota, D.A., Duchek, J.M., Gernbacher, M.A., & Smith, S. (1997). Inhibitory control during sentence comprehension in individuals with dementia of the Alzheimer's type. Brain and Language, 57, 225-253.
4. Houghton, G., Tipper, S.P., Weaver, B., & Shore, D.I. (1996). Inhibition and interference in selective attention: Some tests of a neural network model. Visual Cognition, 3, 119-164.
5. Kanne, S. M., Balota, D. A., Spieler, D. H. & Faust, M. E. (1998). Explorations of Cohen, Dunbar and McClelland's (1990) connectionist model of Stroop performance. Psychological Review, 105, 174-187.

6. MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. Psychological Bulletin, 109, 163-203.
7. Mewhort, D. J. K., Braun, J. G., & Heathcote, A. (1992). Response time distributions and the Stroop task: A test of the Cohen, Dunbar & McClelland (1990) model. Journal of Experimental Psychology, 18, 872-882.
8. Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. Journal of Experimental Psychology, 18, 643-662.