used in the example. The obvious next step would be to find a non-trivial database of dozens or hundreds of propositions to test the model.

Such an effort would require a learning algorithm for IRAAM, which, given a set of propositions or other hierarchical structures, would use gradient descent or a similar method to learn a set of weights encoding just those propositions. Such an algorithm would have an error function assigning a penalty for both missing encodings and for encodings inconsistent with the examples from the training set (e.g., the proposition `female(albert)` in our example data set). An intriguing possibility would be to co-evolve both the network weights and a separate labeling program, using the paradigm described in (Hillis 1992) and (Juillé and Pollack 1996).

Finally, we suspect that inherent limitations in using a single set of network weights may hinder our attempts to use IRAAM as a model of unification in natural language, where the combinatorial possibilities are much richer than those of artificial languages like FOPC and Prolog. Research in image compression (Barnsley and Jacquin 1988) has shown the usefulness of combining several different IFSes to encode a single real-world image. We hope that a similar approach will allow IRAAM to scale up to the larger, more complicated phrases and sentences of natural language.

References

Ackley, D. H., G. Hinton, and T. Sejnowski (1985). A learning algorithm for boltzmann machines. *Cognitive Science 9*, 147–169.

Barnsley, M. F. (1993). *Fractals everywhere*. New York: Academic Press.

Barnsley, M. F. and A. Jacquin (1988). Application of recurrent iterated function systems to images. *In Proc. SPIE*, Volume 1001, pp. 122–131.

Blank, D., L. Meeden, and J. Marshall (1991). Exploring the symbolic/subsymbolic continuum: A case study of raam. Technical Report TR332, Computer Science Department, University of Indiana.

Chalmers, D. (1990). Syntactic transformations on distributed represenations. *Connection Science 2*, 53–62.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Clocksin, W. and C. Mellish (1994). *Programming in Prolog*. Berlin: Springer Verlag.

Dawkins, R. (1986). *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design*. New York: W.W. Norton and Co.

Fodor, J. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J. and Z. Pylyshyn (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition 28*, 3–71.

Hillis, W. (1992). Co-evolving parasites improves simulated evolution as an optimization procedure. In C. Langton, C. Taylor, and J. Farmer (Eds.), *Artificial Life II*, pp. 313–324. Addison Wesley.

Horgan, T. and J. Tienson (1989). Representations without rules. *Philosophical Topics XVII*(1), 147–175.

Juillé, H. and J. Pollack (1996). Co-evolving intertwined spirals. In *Proceedings of the Fifth Annual Conference on Evolutionary Programming*. MIT Press.

Melnik, O. (2000). *Representation of Information in Neural Networks*. Ph. D. thesis, Brandeis University.

Melnik, O., S. Levy, and J. Pollack (2000). Raam for an infinite context-free language. In *IJCNN 2000*. International Joint Conference on Neural Networks: IEEE.

Pinker, S. and A. Prince (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition 28*, 73–193.

Pollack, J. (1990). Recursive distributed representations. *Artifical Intelligence 36*, 77–105.

Rich, E. and K. Knight (1991). *Artificial Intelligence*. New York: McGraw-Hill.

Robinson, J. (1965). A machine-oriented logic based on the resolution principle. *Journal of the ACM 12*(1), 23–41.

Rodriguez, P., J. Wiles, and J. Elman (1999). A recurrent neural network that learns to count. *Connection Science 11*, 5–40.

Rumelhart, D., G. Hinton, and R. Williams (1986). Learning internal representation by error propagation. In D. Rumelhart and J. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1. MIT.

Rumelhart, D. and J. McClelland (1986). On learning the past tenses of english verbs. In D. Rumelhart and J. McClelland (Eds.), *op. cit.*, Volume 2.

Shieber, S. (1986). *An Introduction to Unification Based Approaches to Grammar*. Number 4 in CSLI Lecture Notes. University of Chicago Press.

Siegelmann, H. (1995). Computation beyond the turing limit. *Science 268*, 545–548.

Tabor, W. (2000). Fractal encoding of context-free grammars in connectionist networks. *Expert Systems: The International Journal of Knowledge Engineering and Neural Networks, 17*(1), 41–56.

Van Gelder, T. (1990). Compositionality: a connectionist variation on a classical theme. *Cognitive Science 14*, 355–384.

Williams, R. and D. Zipser (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation 1*, 270–280.

Modeling Selective Attention: Not Just Another Model of Stroop (NJAMOS)

Marsha C. Lovett (Lovett@CMU.EDU)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

The Stroop effect has been studied for more than sixty years, and yet it still defies a complete theoretical account. The model NJAMOS offers a new theoretical account that integrates several explanations of the Stroop phenomenon into a hybrid model. NJAMOS is built within the ACT-R cognitive architecture (Anderson & Lebiere, 1998). Besides fitting a variety of experimental results, NJAMOS offers the potential to capture strategic variation in what is typically considered a low-level attentional phenomenon.

Introduction

The Stroop effect has been studied for more than sixty years (Stroop, 1935), and yet it still defies a complete theoretical account. One explanation for the apparent lack of progress is that so much empirical research has been conducted using this basic paradigm that what we now call the "Stroop effect" is actually a compendium of results derived from a multitude of manipulations applied to a family of Stroop-like tasks! The current article focuses on a select set of Stroop results in order to introduce the model NJAMOS. NJAMOS offers a new theoretical account that integrates several explanations of the Stroop phenomenon into a hybrid model. Specifically, NJAMOS performs competitive, parallel retrieval of information within a goal-based, sequential cognitive processor. NJAMOS is built within the ACT-R cognitive architecture (Anderson & Lebiere, 1998), so it applies a general, pre-specified set of learning and performance mechanisms to the particulars of the Stroop paradigm. Moreover, NJAMOS is unique among models of ("low level") attentional phenomena in that it allows for ("high level") strategic variability.

The organization of the paper is as follows. First a description of the Stroop phenomenon is presented. Then, major theoretical features of other computational models are reviewed. Next, the NJAMOS model is described and fit to a selection of relevant data.

The Basic Phenomenon

The Stroop effect offers a window onto the processes of selective attention in that stimuli with two prominent dimensions are presented in a task where one dimension must be processed and the other ignored. Typically, the stimuli are words, and the two dimensions are the form of the word and the color of the ink in which it is written. The task, then, is either to name the ink color or to read the word. The basic Stroop manipulation varies the relationship between the meaning of the word and the color of the ink to be *congruent* (e.g., the word "red" printed in red ink), *conflicting* (e.g., the word "blue" printed in red ink), or *neutral* (e.g., the word "dog" or a string of "X"s printed in red ink). A robust result emerges: for color naming, there is interference in the conflicting case and (usually) facilitation in the congruent case, but for word reading, there is no (or very little) effect of the congruency of this relationship.

Figure 1 shows a typical data set (along with the NJAMOS predictions to be discussed later). The interference and facilitation in color naming can be seen by the shifts in the color-naming curve as a function of congruency. The lack of such effects for word reading are shown by the relatively flat line for this condition. These results suggest an asymmetry in selective attention, namely, that subjects are strongly influenced by the word when naming the ink color but that they can ignore the ink color when reading.

Theoretical Accounts of the Stroop Effect

Two different views of the Stroop effect cover much of the theoretical work in this area. The "horse-race" view highlights the overall difference in speed of processing for words versus colors (See Figure 1, separation of the two curves) and assumes a response bottleneck. This view implies that the pattern of interference depends on the relative arrival of word versus color information to the response stage: whichever kind of information arrives first will produce interference for the other. Because word reading is, on average, faster than color naming, this view predicts the asymmetry of words interfering with colors but not vice versa.

The other view of the Stroop effect highlights the different levels of automaticity people have acquired for processing the two stimulus dimensions. Because word reading is so highly practiced among typical Stroop experiment subjects, it is more automatic than color naming. This greater automaticity implies that reading requires fewer attentional resources and hence interferes more easily with color naming.

The key similarity between the two views is that they both emphasize parallel processing of the two stimulus dimensions. Not surprisingly, then, the dominant computational accounts of Stroop phenomena have been implemented within connectionist models. The key difference between the two views is whether relative speed or automaticity is considered the main

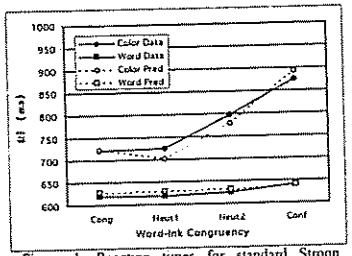

Figure 1. Reaction times for standard Stroop experiment. "Neut1" refers to a string of "X"s in colored ink for color naming and a word printed in black ink for word reading; "neut2" refers to a non-color word in colored ink for color naming.

determiner of interference effects. Note that manipulations of stimulus-onset-asynchrony (SOA) alter the timing of processing predicted to be important under the horse-race view, and manipulations of practice can alter the automaticity of processing that is predicted to be important under the automaticity view. Thus, experiments that employ these manipulations are crucial for evaluating specific theoretical and computational accounts of Stroop phenomena.

Existing Computational Models

Three existing computational models will be discussed. Two of these are connectionist models (Cohen, Dunbar, & McClelland, 1990; Phaf, van der Heijden, & Hudson, 1990), and the other is built as a production system (Roelofs, 2000). The Cohen et al. (1990) model was designed to capture an automaticity account of Stroop. It has input nodes that connect to hidden-layer nodes which connect to response nodes via either a word-reading or color-naming pathway. To represent the greater automaticity of reading, stronger weights are given along the word-reading pathway, making reading less sensitive to the congruency relationship between feature inputs. The Phaf et al. (1990) model was built as the Stroop extension to a general model of selective visual attention. This model's network architecture differs for word reading and color naming in that there are direct input-output connections for word reading only. This removes intermediate (hidden layer) processing for word reading which reduces the potential for semantically based interference in that task.

The Roelofs (2000) model of Stroop phenomena is built as an extension to the WEAVER++ model of word production (Levelt, Roelofs, & Meyer, 1999). As such, this Stroop model specifies separate processing stages for lemma retrieval, word-form encoding, etc. Like the Phaf et al. model, it builds in a word-reading advantage by requiring fewer processing steps, hence enabling different sized congruency effects across the two tasks.

These Models' Fits to Data

These three models all account for the basic results in Figure 1, but how do they fare in predicting other important Stroop results? As mentioned above, one critical Stroop manipulation varies SOA. Glaser and Glaser's (1982) Experiment 1 did this by presenting the two stimulus dimensions spatially separated so that the onset of word and color information could be lagged. SOAs were varied from –400ms to +400ms, where negative SOA means preexposure of the irrelevant information. Figure 2a presents these data.

All three models have simulated these data to varying degrees of success. They all predict the standard Stroop result at 0 SOA, and they all show a reduced Stroop effect for color naming at positive SOAs that is consistent with the data. There are several areas of misfit, however. Both PDP models show a small but notable congruency effect for word reading at negative SOAs, even though this is not present in the data. In addition, the Cohen et al. model predicts that, in color naming, the congruency effect will be largest at negative SOAs, when it is actually reduced here relative to 0 SOA. The Roelofs model has neither of these problems but fails to capture the gradual increase in interference for color naming from –400 to 0 SOA.

The other critical manipulation mentioned above is degree of practice. MacLeod and Dunbar (1988) devised a clever variant of the Stroop task in which they presented stimuli with dimensions shape and color and asked participants to either name the shape or the color. So that congruency of the shape-color relationship could be varied, the shapes came from a fixed set of unfamiliar shapes, and participants were trained to name each shape with a specific color word (e.g., the irregular hexagon is named "red"). Manipulating practice involved giving participants 20 sessions of shape training and measuring their Stroop performance along the way. Figure 3a presents these data.

The Cohen et al. model captures the early (training session 1) pattern of colors interfering with shape naming and not vice versa as well as the late (session 20) pattern of shapes interfering with color naming and hardly vice versa. This reversal is accomplished by the model's weight-learning mechanism that gradually strengthens the shape-naming pathway. An interesting transition point in the data (session 5) shows considerable interference for both tasks. However, the model makes its transition by predicting no interference in either task at this point.

Regarding practice effects, the Phaf et al. and Roelofs models are essentially silent. In both, the word-reading advantage is implemented as an qualitative change (a "short cut" for word-reading), and there is no specific learning mechanism presented.

Summary of Existing Models

To be competitive, any computational model of Stroop effects must demonstrate some added value.

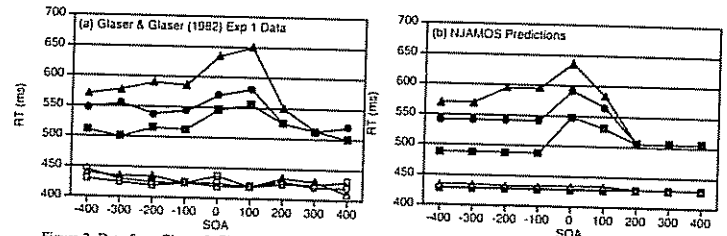

Figure 2. Data from Glaser & Glaser (1982) (a) and NJAMOS predictions (b). RTs plotted for color naming (solid symbols) versus word reading (open symbols) and for conflicting (Δ), neutral (O), and congruent (□) stimuli.

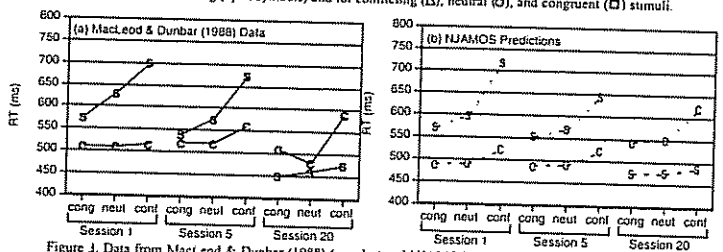

Figure 3. Data from MacLeod & Dunbar (1988) (panel a) and NJAMOS predictions (panel b). In each panel, RTs for shape naming (S) and color naming (C) are plotted separately across congruency and training manipulations.

What do each of these models offer? The Cohen et al. model offers the best coverage of Stroop-related data at present and does so with very simple, natural connectionist mechanisms for strengthening connections and gating activation along the two pathways. This trades off against the model's acknowledged exclusion of other processes (e.g., habituation, strategy choice, etc) that could help address its known areas of misfit.

The Phaf et al. model and the Roelofs model both offer an account of Stroop phenomena within a more general model of a related task. These models demonstrate how processes that play a central role in another task can help explain certain Stroop effects. However, this breadth tends to dilute the models' coverage of the Stroop literature per se.

A Hybrid Model, NJAMOS

There are several sources of added value in NJAMOS. First, it is built within the ACT-R cognitive architecture (Anderson & Lebiere, 1998), so it applies a general, pre-specified set of learning and performance mechanisms to the particulars of the Stroop paradigm. As such, it has many of the benefits of a general approach in that the mechanisms that drive its predictions have been shown to produce accurate predictions for a variety of other tasks. This general-mechanisms approach also offers guidance and constraint in developing the specific task model beyond what the empirical data on that task can provide, i.e., the model must be built to fit the relevant data *within* a given, pre-specified structure (cf. Newell, 1990).

Second, building a specific task model within a cognitive architecture such as ACT-R still allows (and arguably facilitates) a focus on capturing as many experimental results associated with that particular task as possible. Indeed, with the architecture taken as given, model development involves specifying the knowledge used in performing the current task, some of which may come from prior experience and some of which may come from exposure to the task itself. Third, NJAMOS is unique among models of ("low level") attentional phenomena in that it allows for ("high level") strategic variability. Evidence suggests that participants can (and do) apply different strategies when performing the Stroop task (Chen & Johnson, 1991; Logan, Zbrodoff, & Williamson, 1984). This strategic variation may be considered a source of qualitative differences. Note that NJAMOS can also produce quantitative differences in performance by

varying the ACT-R parameter associated with individual differences in working memory capacity (Lovett, Daily, & Reder, 2000).

NJAMOS Model Overview

NJAMOS specifies the knowledge relevant for performing Stroop-like tasks. In ACT-R, this is comprised by a set of production rules and a network of declarative chunks. Each production rule takes the form "IF <conditions> THEN <actions>" and has an associated measure of utility. A production rule's utility is learned through past experiences; it provides an index of how effective the production rule has been. In ACT-R, when more than one production rule's conditions are met, these rules compete on the basis of their utility values. The production rule with the highest utility (after some noise is added) is applied, and its actions are executed. In practice, this means that the system can learn to choose (and prefer) more effective strategies.

A declarative chunk is represented as a template with slots for related information. At the sub-symbolic level, each chunk has a base-level activation representing its overall accessibility. A network of chunks is specified by directed, pairwise connections, each with a strength of association. These strengths influence how strongly one chunk (in the current focus of attention) cues the availability of another chunk. The ACT-R equation for total activation of chunk i is

$$A_i = B_i + \Sigma W_j S_{ji}$$

where B_i is the base-level activation of chunk i, W_j is the amount of attention focused on chunk j, and S_{ji} is the strength of association from chunk j to i. This quantity then translates into the chunk's likelihood of being retrieved and latency to retrieval. The latency measure is more relevant in NJAMOS; it follows the function $L_i = Fe^{-A_i}$ where F is a latency-scale factor.

Some specific examples of the knowledge units specified in NJAMOS (and presented in English for readability) are as follows:

IF the goal involves a color-naming task, and the stimulus color has been encoded, THEN retrieve the associated color concept.

IF the goal involves processing a stimulus [generic goal] and the stimulus has word-like qualities, THEN retrieve the associated word concept.

Color-association-chunk1
Stimulus: /
Feature: color
Concept: black

Current-goal
Stimulus-word: "green"
Stimulus-color: /
Task: color

The first two examples above are production rules that would initiate processing of the color and word dimensions of the stimulus, respectively. The third example is a declarative chunk representing information about the color black. The last example is a specification of a goal chunk for the color-naming task after the stimulus has been perceived but before any color concept has been retrieved (either by processing the word or the color). In ACT-R, processing is in large part driven by the current goal. The goal is represented like other declarative chunks but it is considered to be the current focus of attention. A fixed, limited amount of source activation (W in ACT-R) is shared among the different slots of the current goal (W_j for each slot). This share of source activation is spread to connected chunks (proportionally to the strength of the connection) and added to the receiving chunk's base-level activation to comprise that chunk's total activation.

Given this specification of knowledge and ACT-R's fixed mechanisms, two important features of NJAMOS's performance follow. First, production rule choice favors word reading (even when the task is color naming). This is because the word-processing production (see above) is general enough that it applies whenever there is a word-like stimulus and because its utility value is taken to be much higher than that for color naming. (The former feature is taken to represent the extensive practice of reading such that words are processed regardless of the current goal; the latter comes from ACT-R's utility-learning mechanism given this extensive practice.[1]) Production-rule choice provides NJAMOS with a degree of sequentiality as each production rule in this model tends to focus on a single dimension (word or color). This also provides NJAMOS with strategic variability: processing of the word may either precede or follow processing of the color. Either way, a "check" production works to verify that the correct dimension was processed before enabling a response.

Parallel processing plays a role in producing the second important feature of NJAMOS performance, namely, that both dimensions of the stimulus influence the retrieval of declarative chunks. This is because all of the slots in the current goal provide contextual cues to influence the total activation of to-be-retrieved chunks (see total activation equation). Some of these slots represent the stimulus features (word *and* color); others may represent a concept retrieved during previous processing (e.g., if the word were partially processed when the word-processing production won over a color-processing production, then the concept of the word would provide another cue when retrieving the color of the ink). Recall that these contextual effects are proportional to the strengths of association between the relevant chunks. In NJAMOS, pairs of chunks that involve the same color concept (e.g., a color-association chunk for blue and a word-association chunk for blue) are given positive strengths of association, and chunks that involve different color concepts have negative strengths of association. These positive and negative strengths contribute to the facilitation and interference effects. NJAMOS produces larger facilitation and interference effects in color naming (see Figure 1) predominantly because (a) color-association chunks are given a lower base-level activation which, in the nonlinear latency equation (see above), makes them not only slower overall but more sensitive to contextual cueing and (b) the concept of the word will often have been retrieved into the goal before retrieving the color (but not vice versa), magnifying the contextual effect of the word dimension of the stimulus. Figure 4 presents a sketch of the flow of control in NJAMOS, summarizing these relationships and indicating other examples of production-rule choice and parallel chunk retrievals.

[1] Note that the model's performance does not require that color naming be constrained by a specific goal. As long as word reading has a higher utility, both word-reading and color-naming productions could fire under a generic goal.

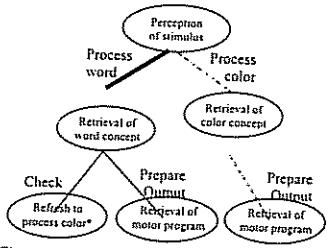

Figure 4. Flow of control in NJAMOS for color naming with standard stimuli. Lines represent production rules (note choices), and ovals represent parallel retrievals that modify the current focus of attention. From the bottom left oval (*), flow continues along the dashed path.

Fitting NJAMOS to the Data

This section describes how NJAMOS was fit to five separate experiments comprising a total of 92 data points (including Figure 1). For ease in estimating a set of best-fitting parameters, a mathematical description of NJAMOS was used. The same parameter settings were used to predict performance across all five experiments[2] with the following exceptions. For each experiment two parameters were varied. One of these is the latency-scale factor F (see latency equation above), and the other was a latency intercept. These parameters appeared necessary to capture the experiment-to-experiment variation in latencies that appear even under very similar manipulations and experimental designs (e.g., compare RTs in Figures 1 and 5). In a couple cases, an additional single parameter was varied when relevant to a particular experiment, and those cases will be discussed below. To preview the quality of the NJAMOS fit, all 92 data points were fit by varying a total of 12 parameters for a R^2 of .95 and MSE of 1114. In essence, the goal was to test whether NJAMOS can produce reasonable accounts of these experiments without very much parameter varying.

[2] Parameters taken as fixed but whose values departed from default include base-level activation for word-association chunks (set at 2) and for color-association chunks (set at 0), S_{ji}'s between same-color chunks (set at 1.5), between different-color chunks (set at −1.5), between same-task chunks (set at 0.6), & between different-task chunks (set at 0).

Dunbar & MacLeod (1984)

Figure 5 shows the data and NJAMOS predictions for Dunbar and MacLeod's experiment (1984, p. 62). The design is similar to that presented in Figure 1, with the neutral condition corresponding to "Neut1." Both the data and model show the standard Stroop results.

Glaser & Glaser (1982)

Figure 2b shows the predictions of NJAMOS for the Glaser and Glaser (1982) experiment described above. Unlike previous models, NJAMOS appears to capture (at least qualitatively) three effects in the data: (1) no congruency effect for word reading at any SOA, (2) increase in overall latency for color naming as SOA goes from −400 to 0, and (3) increase and then decrease in interference effects from −400 to 0 to + 400 SOA. Much of this match to the data is driven by the model's tendency to choose word reading first, even on color naming trials. Note, however, that the size of the interference effect is not monotonically increasing from −400 to 0 as it is in the data and the facilitation effect in color naming is overpredicted. Under the parameter-fitting constraints applied here, this is not too surprising. The fit can be improved by allowing more parameters to vary (e.g., modifying one or two S_{ji} values has an impact).

MacLeod & Dunbar (1988)

Figure 3b presents the NJAMOS predictions for Dunbar and MacLeod's (1988) experiment. Note that this is a case where an additional parameter was required by the design of the experiment, namely, the initial base-level activation for the shape-name-association chunks. From there, ACT-R's learning mechanism naturally increased this base-level activation as the model underwent training by getting additional practice retrievals of the shape-name-association chunks. This makes the shape chunks less susceptible to context effects. Also, NJAMOS can learn to adjust its strategy of trying to name the color first (regardless of task) as the two competing (shape and color-processing) productions adjust their utilities. This changes the degree to which color information is present in the goal to add to the

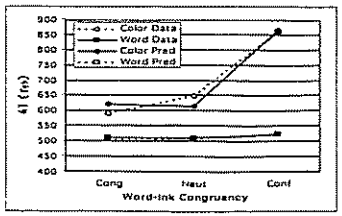

Figure 5. Dunbar & MacLeod (1984) data and model.

facilitating/interfering effect. In this model fit, procedural learning played a relatively small role.

NJAMOS shows three main results in this experiment: (1) at session 1, there is more interference and facilitation in shape naming than in color naming, (2) at session 5, there is some facilitation and interference in both tasks, and at session 20, there is more facilitation and interference in color naming than shape naming. Yet, the data show an increase in the interference effect for color naming between sessions 1 and 5 that is hardly present in the predictions. Also, the data show almost no interference of shapes on color naming in session 1 (5ms), whereas the model shows more (30ms). It is possible that adjusting the production parameters would address both of these issues, but for current purposes the goal was to find a reasonable model fit with a small number of free parameters.

Cohen et al. (1992)

Table 1 presents Stroop data from schizophrenic patients and matched controls along with NJAMOS predictions. The main result is a larger interference effect (for color naming) among the patients. While schizophrenia is a very complex condition, it is associated with a deficit in working memory. To capture this NJAMOS simply uses an adjusted W parameter. This was the extra parameter varied to fit these data. As the table shows, reducing W produces the effect as well as raising RTs among the patients.

Table 1. Data and model fit to schizophrenic data

	Neut-Word	Neut-Color	Conf-Color
Patient data	530	797	1467
Control data	420	603	1037
Patient pred	456	764	1345
Control pred	440	661	1137

Conclusions

The above model fits demonstrate that NJAMOS can fit a variety of Stroop results, when constrained by the ACT-R architecture and by limited free parameters. What does this mean for the value of the model? In the case of Stroop, the proof of a model comes from many sources: parsimony, generality, and coverage of many results. NJAMOS aspires for all these. Because of the breadth of Stroop results, however, any model will likely focus on a subset. MacLeod's (1991) review lists 18 key results that may serve as a core, but even then, models can differ in how directly they address these various results. The best course may be to use Stroop models as a theoretical tool for understanding not only how well but *why* a certain result is or is not captured.

References

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought.* Mahwah, NJ: Erlbaum.

Chen, J., & Johnson, M. (1991). The Stroop congruency effect is more observable under a speed strategy than an accuracy strategy. *Perc. & Motor Skills, 73,* 67-76.

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.

Cohen, J. D., Servan-Schreiber, D., & McClelland, J. (1992). A parallel distributed processing approach to automaticity. *Am J of Psych,* 239-269.

Dunbar, K., & MacLeod, C. M. (1984). A horse race of a different color: Stroop interference patterns with transformed words. *JEP: HPP, 10,* 622-639.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *JEP:HPP, 8,* 875-894.

Levelt, W. J. M., Roelofs, A., Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral & Brain Science, 22,* 1-75.

Logan, G. D., Zbrodoff, N. J., Williamson, J. (1984). Strategies in the color-word Stroop task. *Bulletin of the Psychonomic Society , 22,* 135-138.

Lovett, M. C., Daily, L. Z., & Reder, L. M. (2000). A source activation theory of working memory. *Cognitive Systems Research, 1,* 99-118.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163-203.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 126-135.

Newell, A. (1990). *Unified Theories of Cognition.* Cambridge, MA: Harvard University Press.

Phaf, R. H., van der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology, 22,* 273-341.

Roelofs, A. (2000). Control of language: A computational account of the Stroop asymmetry. *Proc. of the Third Int'l Conf on Cognitive Modeling.* Veenendaal, The Netherlands: Universal Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643-662.

Learning of Joint Visual Attention by Reinforcement Learning

Goh Matsuda
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo, 153-8902 Japan

Takashi Omori
Graduate School of Engineering, Hokkaido University
Kita-ku, Kita 13 Jou Nishi 8 Cho, Sapporo, 060-8628 Japan

Abstract

In this paper, we propose a neural network model of joint visual attention learning that plays an important role in infant development, and we discuss previous studies of experimental psychology on joint visual attention based on simulation results using the model. We assumed an imaginary experiment to develop the model. A mother and an infant are sitting face to face with a table between them. Some objects familiar to the infant are placed on the table, and toys operated by remote control are put outside of the infant's view. The infant is given a reward of seeing something interesting only when the infant follows the mother's gaze after eye contact. We constructed the model of this experiment with a reinforcement learning algorithm, and simulated the experiment on a computer. As a result, it was revealed that the infant could learn a series of joint-visual-attention-like actions by receiving rewards from an environment, although it initially has little knowledge of the environment. This result suggests that infants can acquire joint visual attention without comprehension of the nature of joint attention, i.e., "I'm looking at the same thing that others are looking at."

Introduction

Modeling the development of infant intelligence is one of the strategies for understanding human intelligence. We focus on development in infancy from the viewpoint of engineering. Neonatal babies have little knowledge of the environment, nevertheless they acquire new knowledge and behavior suitable for the environment step by step in their developmental stage. Although the whole brain system of adults is very complicated, we believe that we can create a model of intelligence relatively easily by pursuing those developmental steps one by one. In this study, we focus on joint visual attention as one of those developmental processes.

In an engineering sense, joint attention can be defined as the sharing of attention with others, and joint visual attention is defined as looking at what others are looking at. Although this definition may cause some objections, we adopt it in this paper.

The detailed study of joint visual attention began with Scaife and Bruner's work (1975). They observed that a child in early and middle infancy follows an adult's gaze, and stated that this behavior is an important factor in early development. However, it is not yet clear how we acquire joint visual attention. There are two theories, nature and nurture, at present (Baron-Cohen, 1995; Butterworth & Jarrett, 1991; Corkum & Moore, 1995).

In this paper, we propose an engineering model of joint visual attention learning by conditioning with signals from the environment, and examine this behavior by computer simulation. Based on the results, we discuss the requirements of fundamental parts that are necessary for such learning.

Behavior Acquisition by Reinforcement Learning

Imaginary Experiment

We contrived the following imaginary experiment for our study based on the behavioral experiment of Corkum and Moore (1995). A mother and an infant are sitting face to face with a table between them. Some objects familiar to the infant are placed on the table. In the early stage of learning, the infant randomly directs its attention to the objects including the mother's face. The mother, however, is always gazing at the infant's eyes. Toys are set outside of the infant's view, and the observer can operate them by remote control (Figure 1). When the infant looks at the mother's face and they make eye contact, the mother moves her eyes to gaze toward any one of the toys. Furthermore, when the infant follows the direction of her gaze , the observer operates the toy, arousing pleasure in the infant's mind. In other words, the infant can obtain a reward of seeing something interesting only when it perform the two consecutive actions of looking at the mother's face and following her gaze.

Temporal Difference Learning

In this study, we used the temporal difference (TD) method (Sutton & Barto, 1998) for the learning of joint visual attention. TD learning is an algorithm that learns the value function $V(s_t)$ of each state s_t based on a reward r that is given later from the environment. An agent learns behavior strategy so that the value function may increase ($0 < \gamma < 1$).