The Time Course of Information Integration in Sentence Processing

Michael J. Spivey, Stanka A. Fitneva
Cornell University

Whitney Tabor
University of Connecticut

Sameer Ajmani
MIT

Abstract

Recent work in sentence processing has highlighted the distinction between serial and parallel application of linguistic constraints in real time. In looking at context effects in syntactic ambiguity resolution, some studies have reported an immediate influence of semantic and discourse information on syntactic parsing (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998; Spivey & Tanenhaus, 1998). However, in looking at the effects of various constraints on grammaticality judgments, some studies have reported a temporal precedence of structural information over semantic information (e.g., McElree & Griffith, 1995, 1998). This chapter points to some computational demonstrations of how an apparent temporal dissociation between structural and non-structural information can in fact arise from the dynamics of the processing system, rather than from its architecture, coupled with the specific parameters of the individual stimuli. A prediction of parallel competitive processing systems is then empirically tested with a new methodology: speeded sentence completions. Results are consistent with a parallel account of the application of linguistic constraints and a competitive account of ambiguity resolution.

Introduction

For more than a couple of decades now many psycholinguists have been investing a great deal of effort into elucidating the "sequence of stages" involved in the comprehension of language. Emphasis has been placed on the question: When do different information sources (syntax, semantics, etc.) get extracted from the linguistic input? One answer to this question that has been very influential is that the computation of syntax precedes the computation of semantics and pragmatics (e.g., Frazier & Fodor, 1978; Ferreira & Clifton, 1986; McElree & Griffith, 1995, 1998). One opposing answer that is gaining support is that there are no architecturally imposed delays of information during sentence processing, that all relevant information sources are extracted and used the moment they are received as input (MacDonald, Pearlmutter, & Seidenberg, 1994; Spivey-Knowlton & Sedivy, 1995; Trueswell & Tanenhaus, 1994). Recently, however, some disillusionment has been expressed concerning the question itself:

> "Given the wide range of results that have been reported, it seems most appropriate at the moment to determine the situations in which context does and does not have an influence on parsing, rather than continue the debate of *when* context has its impact." (Clifton, Frazier, & Rayner, 1994, p.10, italics theirs).

Perhaps one way to redirect the "when" question to better understand the mixed results in the literature would be to turn it into a "how" question. Could *the manner in which* various information sources combine during sentence processing wind up explaining why context sometimes has an early influence and sometimes a late influence? It seems clear that a treatment of this kind of question will require some theoretical constructs and experimental methodologies that are new to sentence processing, as well as some careful attention to lexically-specific variation in stimulus items. The purpose of this chapter is to describe some of these new approaches and the implications that they have for claims about the time course of information integration in sentence processing.

Nonlinear Dynamics

Over the past fifteen years, a number of researchers have designed dynamical models of sentence processing (Cottrell & Small, 1983; Elman, 1991; McClelland & Kawamoto, 1986; McRae, Spivey-Knowlton & Tanenhaus, 1998; Selman & Hirst, 1985; Spivey & Tanenhaus, 1998; St. John & McClelland, 1990; Tabor & Hutchins, 2000; Tabor, Juliano, & Tanenhaus, 1997; Waltz & Pollack, 1985; Wiles & Elman, 1995; see also Henderson, 1994, and Stevenson, 1993, for hybrid models that combine rule-based systems with some fine-grain temporal dynamics). A dynamical model is a formal model that can be described in terms of how it changes. Typically, such models take the form of a differential equation,

$$d\mathbf{x}/dt = f(\mathbf{x}) \qquad \text{(Eq. 1)}$$

with an initial condition, $\mathbf{x} = \mathbf{x}_0$. Here x is a vector of several dimensions and t is time. The equation says that the change in $\mathbf{x}$ can be computed from the current value of $\mathbf{x}$. The behavior of such systems is often organized around *attractors*, or stable states ($f(\mathbf{x})$=0) that the system goes toward from nearby positions. Nearby attractors will tend to have a strong "gravitational pull", and more distant attractors will have a weaker pull. The most common strategy is to assume that initial conditions are determined by the current context (e.g., a string of words like "Alison ran the coffee-grinder") and that attractors correspond to interpretations of that context (e.g. Alison is the agent of a machine-operation event where the machine is a coffee-grinder). The model, (Eq. 1), is called *nonlinear* if f is a nonlinear function. Nonlinearity is a necessary consequence of having more than one attractor. Since languages contain many sentences with different interpretations (and many partial sentences with different partial interpretations), dynamical models of sentence processing are usually highly nonlinear. The potential for feedback in Equation (1) -- the current value of a particular dimension of **x** can depend on its past value -- is also important. It can cause the system to vacillate in a complex manner before settling into an attractor.

Many dynamical sentence processing models are implemented in connectionist models (i.e., artificial neural networks). The "neural" activation values correspond to the dimensions of the vector **x** and the activation update rules correspond (implicitly) to the function, f. In some such cases (e.g., Elman, 1991; St. John & McClelland, 1990; Wiles & Elman, 1995), Equation (1) is replaced by an iterated mapping (Eq. 2):

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t) \qquad \text{(Eq. 2)}$$

which makes large discrete, rather than continuous, or approximately continuous, changes in the activation values. Typically, such discrete models are designed so that words are presented to the model one at a time and activation flows in a feedforward manner upon presentation of a single word. This architecture makes no use of the feedback potential of Equation (1), so the dynamics of single word-presentations are trivial; but over the course of several word presentations, activation can flow in circuits around the network, and feedback (as well as input) can contribute significantly to the complexity of the trajectories (Wiles & Elman, 1995). Other proposals allow feedback to cycle after every input presentation. Some such proposals present all the words in a sentence at once (Selman & Hirst, 1985), while others use serial word presentation and allow cycling after each word (Cottrell & Small, 1983; McRae et al., 1998; Spivey & Tanenhaus, 1998; Tabor & Hutchins, 2000; Tabor et al., 1997; Waltz & Pollack, 1985; Wiles & Elman, 1995).

Models which allow feedback to cycle after each input make fine-grained predictions about the time course of information integration in sentence processing. In fact, several existing dynamical models of sentence processing exhibit at least simple forms of vacillation. For example, when presented with the string, "Bob threw up dinner", Cottrell and Small (1983)'s model shows a node corresponding to the *purposely propel* sense of "throw" first gaining and then losing activation (see also Kawamoto, 1993). Tabor et al. (1997) define a dynamical system in which isolated stable states correspond to partial parses of partial strings. At the word "the" in the partial sentence, "A woman insisted the...", for example, they observe a trajectory which curves first toward and then away from an attractor corresponding to the (grammatically impossible) hypothesis that "the" is the determiner of a direct object of "insisted", before reaching an (grammatically appropriate) attractor corresponding to the hypothesis that "the" is the determiner of the subject of an embedded clause. Syntax-first models of sentence processing (Frazier & Fodor, 1978; Frazier, 1987; McElree & Griffith, 1998) are typically designed to restrict vacillation to a very simple form: first one constraint system (syntax) chooses a parse instantaneously and then another one (e.g., semantics) revises it if necessary.

In *lexical* ambiguity resolution, there is evidence for another simple form of vacillation. Tanenhaus, Leiman, and Seidenberg (1979, see also Swinney, 1979, and Kawamoto, 1993), found that ambiguous words exhibit temporary (approx. 200 ms) priming of both meanings (e.g. "rose" as *flower* and "rose" as *moved up*) even in a context where only one meaning is appropriate (e.g. "She held the rose"). Soon thereafter, the contextually inappropriate meaning ceases to exhibit priming. Recent constraint-based models of parsing predict effects in syntactic ambiguity resolution that significantly resemble the effects in lexical ambiguity resolution (MacDonald et al., 1994; Spivey & Tanenhaus, 1998; Trueswell & Tanenhaus, 1994). In contrast, typical syntax-first models of sentence processing posit syntactic parsing strategies that immediately select a single structural alternative (Frazier & Fodor, 1978; Frazier, 1987). To test these two types of models, what we need are experimental methodologies that provide access to the moment-by-moment representations computed during syntactic parsing. Do we see early vacillation between syntactic alternatives, as is seen between lexical alternatives? In this chapter, we will discuss two experimental methodologies that show promise for revealing the temporal dynamics of syntax-related information during sentence processing: speeded grammaticality judgments (McElree & Griffith, 1995, 1998), and speeded sentence completions. Results from these methodologies are simulated by a nonlinear competition algorithm called Normalized Recurrence (Filip, Tanenhaus, Carlson, Allopenna, & Blatt, this volume; McRae et al., 1998; Spivey & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, & Hanna, 2000).

Normalized Recurrence is a relatively simple dynamical system in which the alternative interpretations that a given stimulus might map onto are treated as localist units in the network. The multiple information sources that might give evidence for these different interpretations are then given localist units representing their support for the various stimulus-interpretation mappings. See Figure 1. First, each of the information sources has their previous activations normalized to a sum of 1.0:

$$S_{c,a}(t)=S_{c,a}(t-1) / \sum_{a} S_{c,a}(t-1) \qquad \text{(Eq. 3)}$$

where $S_{c,a}(t)$ is the activation of the *c*th information source supporting the *a*th alternative at time *t*. Next, the information sources combine in a weighted sum at the interpretation units:

$$I_a(t) = \sum_c [w_c * S_{c,a}(t)] \qquad \text{(Eq. 4)}$$

where $I_a(t)$ is the activation of the *a*th alternative interpretation at time *t*, and the weights, w_c -- one for each information source -- sum to 1.0. When an interpretation unit reaches a criterion activation, some appropriate output is stochastically triggered, such that the activation function across the different interpretation units is treated as a probability density function describing the likelihood of each interpretation triggering its preferred action (e.g., looking at the object corresponding to that interpretation, Spivey-Knowlton & Allopenna, 1997). The final computation that completes a cycle of competition is feedback from the integration units to the information sources, where an information source's weighted activations are scaled by the resulting interpretation node's activation and sent as cumulative feedback to the information source (Eq. 5). This feedback is how the model gradually approaches a stable state, coercing not only the interpretation units to settle on one alternative, but also coercing the information sources to conform.

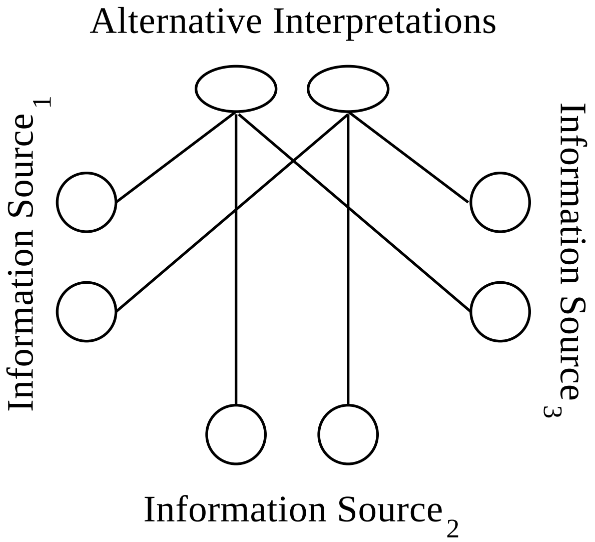

Figure 1. A schematic of the Normalized Recurrence competition algorithm with three information sources competing over two alternatives.

$$S_{c,a}(t+1) = S_{c,a}(t) + I_a(t)*w_c*S_{c,a}(t) \qquad \text{(Eq. 5)}$$

It should be noted that, unlike many connectionist models, this network does not "learn" its weights. Instead, they are each set to 1/*n*(where *n* is the number of information sources, Spivey & Tanenhaus, 1998), or the entire weight space is sampled and the weights with the best fit to the data are used. For example, McRae et al. (1998) designed a Normalized Recurrence network to simulate sentence completion data and self-paced reading data on the Reduced Relative/Main Clause ambiguity. Initially combining three information sources (a general main-clause bias, thematic fit information, and verb tense frequency), and sampling the entire range of weights, it was found that the best weights for fitting that data set were the following: main-clause bias =.5094, thematic fit =.3684, and verb tense frequency =.1222. However, with different stimulus sets and different presentation circumstances that emphasize their information sources differently, the weights for these constraints are likely to vary somewhat.

Highly simplified in comparison to attractor networks that use distributed representations (e.g., Tabor et al., 1997), Normalized Recurrence thereby allows an easily interpreted "peek" into the system's state at any point in time. Panels A and B of Figure 2 show some generic examples of the activation of two alternative interpretations competing over time. Nonlinear trajectories through the state-space on the way toward settling on one alternative canproduce complex behavior in the model. In fact, when several information sources compete over three or more interpretations, an alternative whose initial activation starts out in "second place" can sometimes wind up usurping the most active alternative and eventually become the final interpretation (Figure 2C).

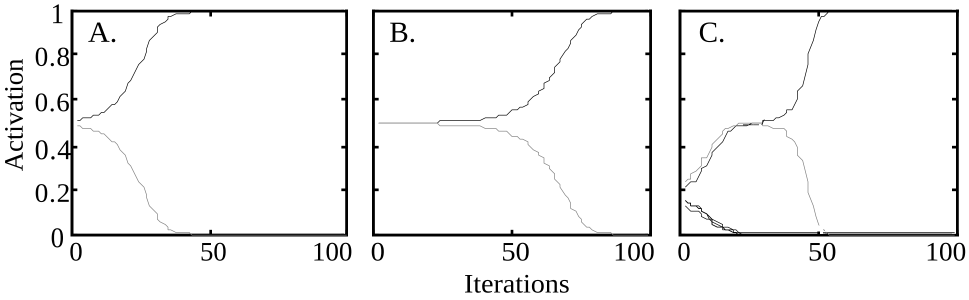

Figure 2. Example results from Normalized Recurrence. Panels A and B are from a network with an architecture like that in Figure 1. Panel C is from a network with four information sources competing over six alternatives. Note that the alternative that starts out with the highest activation (dashed line) ends up losing.

Measures of the Activation of Linguistic Representations

While modeling allows a kind of "x-ray vision" into the internal working parts of a system that might be functioning in a fashion similar to that of the mind, psycholinguists are typically more interested in getting that kind of "x-ray vision" for the *actual* mind -- not an idealized set of formulas intended to simulate the mind. To this end, a number of experimental methodologies have been used over the past couple of decades to tap into the salience of certain linguistic representations *during real-time language processing*. Most of them have been using differences in reaction times to infer relative activations of linguistic representations. It is assumed that a faster reaction time implies a representation with some unspecified amount of greater activation. Although this assumption seems fair enough, determining the mapping from latencies to activations has been largely ignored. What would be preferable would be to see experimental data reflecting the activation of a linguistic representation changing over time, much like those in Figure 2.

One recent example of this kind of "window" into the moment-by-moment activation of different linguistic representations is research with headband-mounted eyetracking (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In experiments looking at spoken word recognition, it was observed that when participants were instructed to "pick up the candy" they tended to briefly fixate a *candle* before finally fixating and grasping the candy. In fact, plotting the probability of fixating the various objects across time produced curves that were surprisingly close to the lexical activation functions from the TRACE model of speech perception (McClelland & Elman, 1986) -- not unlike those in Figure 2C.

Another recent example that shows similar time-slices in the temporal dynamics of linguistic representations is McElree and Griffith's (1995, 1998) use of the speed-accuracy trade-off (SAT) procedure with speeded grammaticality judgments. When the last word in a sentence makes it grammatical or ungrammatical, a rushed decision on this grammaticality is likely to be based on only partially complete representations. By applying signal detection theory to these rushed decisions over various time intervals, McElree and Griffith show a smooth, gradual increase in the detectability of the grammaticality over time -- as measured by d-prime, which provides an index of a subject's sensitivity to a stimulus irrespective of his/her response criteria. In the following sections of this chapter, we will review some of McElree and Griffith's findings and conclusions, test the Normalized Recurrence competition algorithm on their results, as well as introduce some results from a new speeded response methodology: speeded sentence completions. We wish to illustrate how, with the recurrent interplay between experimental data and model simulations, we can iteratively refine a sound theory of the time course of information integration in sentence processing.

Serial Stages in Sentence Processing

The first question that arises in understanding how a serial system might work is the size of the unit of computation. In this kind of treatment, a particular processing stage does not send output to the next stage until it has received (and performed its operations on) an entire unit of computation. In the case of sentence processing, a number of proposals have been forwarded for the size of such units. The temporally-extended unit of serial computation has been suggested to be as large as entire clauses (Fodor, Bever & Garret, 1974) or as small as individual words (Frazier & Fodor, 1978). Alternatively, the serial system could be smoothly cascading, but have a kind of "raw transmission time" between modules (McClelland, 1979). For example, McElree and Griffith (1995) have postulated a ~100 ms delay between the initial computation of subcategory information and the initial computation of thematic role information. More recent work has suggested a 200-400ms delay between syntactic information and lexical information (McElree & Griffith, 1998).

McElree and Griffith's SAT analysis of speeded grammaticality judgments is particularly exciting in that it provides a glimpse into the activation of certain linguistic representational formats (syntax, thematic roles, subcategory constraints, etc.) in real time. In this task, subjects are presented grammatical and ungrammatical sentences, and instructed to, as quickly as possible, judge their grammaticality. As our interest is in *when* various information sources begin to affect the grammaticality judgment, our primary focus will be on the *ungrammatical* sentences. According to McElree and Griffith, sentences like (1a) become ungrammatical at the final word due to a subcategorization violation, because the verb *agreed* is intransitive. In contrast, sentences like (2a) become ungrammatical at the final word due to a thematic role violation, because the Agent of the verb *loved* must be animate (and books are inanimate). In order to compute d-primes via signal detection theory (Green & Swets, 1966) for the SAT task, the ungrammatical sentences (1a & 2a) provided the signal+noise trials and the grammatical sentences (1b & 1b) provided the noise trials. (Thus, the SAT analysis actually treats the task as one of "ungrammaticality detection", rather than grammaticality judgment.)

(1) a. Some people were agreed by books. (Subcategory Violation Sentence)
b. Some people were agreed with rarely. (Subcategory Control Sentence)

(2) a. Some people were loved by books. (Thematic Violation Sentence)
b. Some books were loved by people. (Thematic Control Sentence)

In the SAT version of this speeded grammaticality judgment task, the target sentences were presented to subjects one word at a time in the center of the screen in a noncumulative fashion. Immediately, or shortly, after presentation of the last word in the sentence, a tone would signal to the subject that she/he must respond as to the grammaticality of the sentence within 300 ms. The temporal interval between the onset of the last word and the presentation of the tone was either, 14, 157, 300, 557, 800, 1500, or 3000 ms. (After a couple hours of practice, subjects eventually became skilled at forcing themselves to respond within 300 ms of the tone, even though their processing of the sentence, at the very short intervals, was incomplete.) As seen in Figure 3, mean d-prime values (across six subjects) at the shortest intervals were at or near chance performance. However, at the intermediate and later intervals, performance clearly improved in a smooth, graded fashion. Interestingly, detection of ungrammatical sentences was slightly better for subcategory violations (filled circles) than for thematic role violations (open circles).

One possible interpretation of the data in Figure 5 is that they come from two different exponential functions, each with its own x-intercept. For example, if one extended the left hand portions of the two curves in the simple downward direction implied by the data points at those first few intervals, they would reach a d-prime of zero at slightly different places along the horizontal time axis. If one assumes a dual-process serial processing system, one could infer from these different x-intercepts (as long as the variability in processing

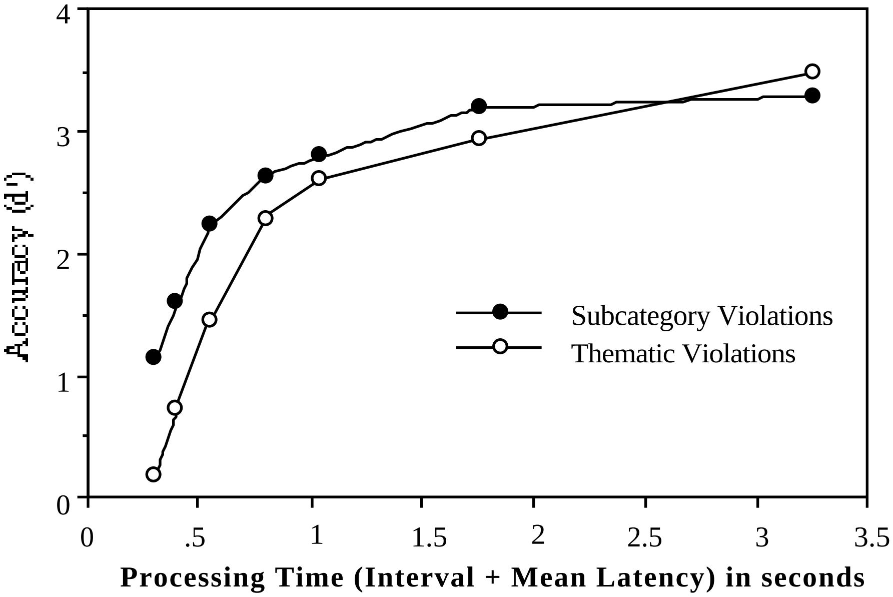

Figure 3. Accuracy of grammaticality detection for subcategory and thematic violations. (Adapted from McElree & Griffith, 1995.)

time is equal across conditions) that subcategorization information "becomes operative", and informs the detection of ungrammaticality, about 100 ms before thematic role information does. In fact, using an exponential equation (Eq. 6) to fit the data points, McElree and Griffith (1995) suggest exactly that.

$$d'(t) = \lambda(1-e^{-\beta(t-\delta)}), \text{ for } t>\delta \text{, else } 0 \qquad \text{(Eq. 6)}$$

In Equation 6, accuracy (*d'*) at each fraction of a second *t* is determined by three free parameters: λ, β, and δ. As the scalar of the entire equation, λ determines the asymptote of the curve, where improvement in accuracy over time tapers off and total accuracy "maxes out". As the scalar in the exponent of *e*, β determines the rate of rise in d-prime over time, or the slope of the curve as it departs from zero. Finally, as the time relative (because it is subtracted from *t*) portion of the exponent of *e*, δ determines the x-intercept of the curve, or the point in time immediately before accuracy climbs above chance. Thus, at the point in time where the curve is to reach zero, *t* and δ will be equal to one another, and *t*-δ will equal zero, making the entire equation equal zero.

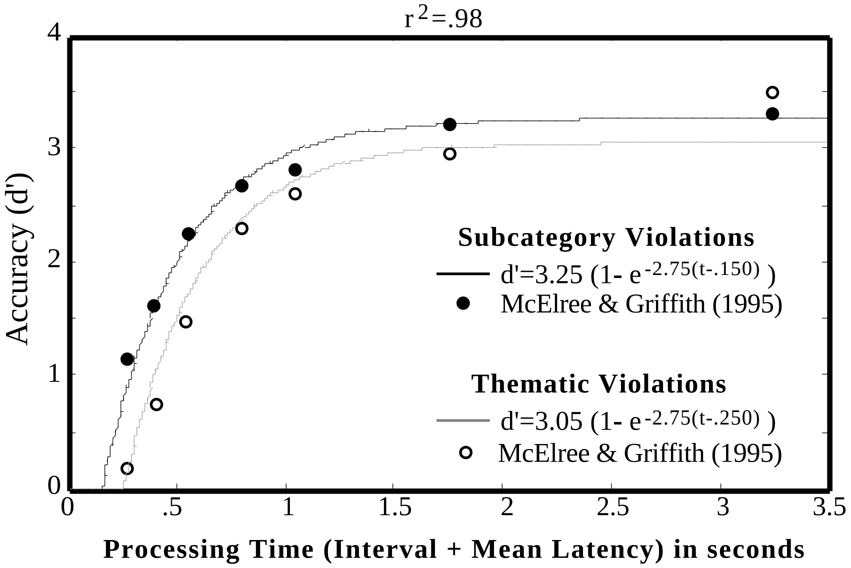

Figure 4. Accuracy of grammaticality detection and approximated fits from McElree and Griffith's (1995) dual-process serial processing account: Equation 6. The fit to the data accounts for 98% of the variance. (Adapted from McElree & Griffith, 1995.)

As negative d-primes would imply a perverse pattern in the data, the last part of the equation insures that for values of *t* that would produce negative d-primes, d-prime is instead rectified to zero. To fit these parameters to the data curves, McElree and Griffith apply Chandler's (1967) Stepit algorithm that searches the parameter space to find the best-fitting parameter values -- somewhat similar to that carried out by McRae et al. (1998) in setting the weights for the Normalized Recurrence competition algorithm. Figure 4 shows an example of the data being fit by the equation, using different δ values (and different λ values) for subcategory violations and thematic violations.

Importantly, this equation provides a standardized method of estimating where the d-prime curves over time would reach zero if they had been sampled from an exponential function that actually had an x-intercept. However, it is certainly possible, in principle, that the data points in Figure 3 do not come from a function with a real x-intercept, but instead come from a function that never actually touches the x-axis, such as the logistic in Figure 5. With no actual x-intercepts (instead, each curve's y-intercept signifies a nonzero d' at timestep 1 -- and is rectified to zero at timestep zero, similar to the rectification done in McElree and Griffith's equation), it would be impossible to make any claims about separate processes "becoming operative" at different discrete points in time.

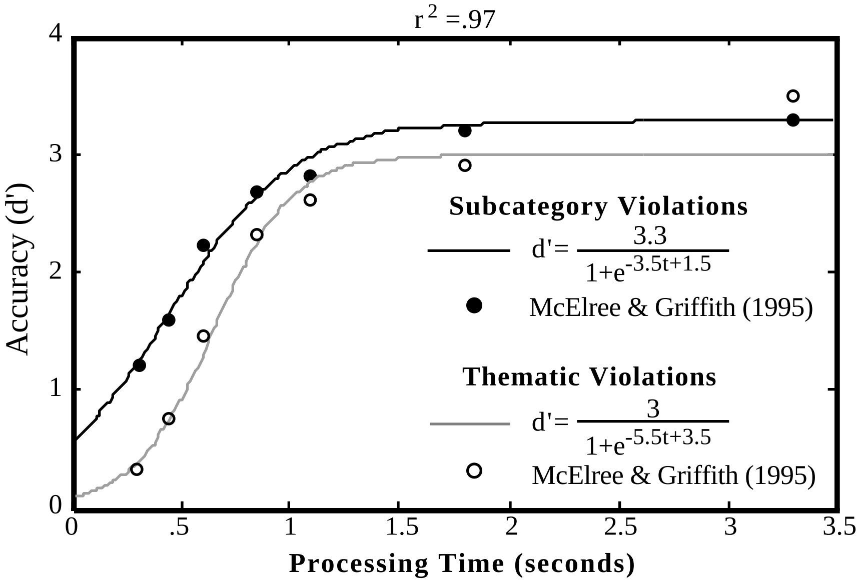

Figure 5. Accuracy of grammaticality detection and approximated fits from a logistic function. The fit to the data accounts for 97% of the variance.

Parallel Integration in Sentence Processing

McElree and Griffith (1995) anticipated that, far from requiring a serial-stage account of sentence processing, their results might in fact be accommodated by certain parallel models of information processing. As the sigmoidal function in Figure 5 is a natural result of competition in Normalized Recurrence, we decided to test Normalized Recurrence on McElree and Griffith's results. To apply the Normalized Recurrence competition algorithm to this grammaticality judgment task, the two information sources (subcategorization and thematic roles) were each condensed into two values: one for the probability of the sentence being grammatical, and one for the probability of the sentence being ungrammatical, based on that information source's strength of constraint. Thus, rather than becoming operative at an earlier point in time, subcategorization information may simply provide a probabilistically stronger constraint on grammaticality than thematic role information does. That is, it may be the case that thematic fit is more violable in our typical language experience (e.g., "This computer hates me.") than subcategorization constraints (e.g., "I slept the day away."). Figure 6 shows a schematic diagram of the Normalized Recurrence model, with bidirectional connections between the information sources and the integration layer (where grammaticality judgment takes place) allowing converging/conflicting biases to be passed back and forth.

As in other Normalized Recurrence simulations, competition between mutually exclusive representations ("grammatical" and "ungrammatical", in this case) proceeded with three critical steps for each iteration of the model: 1) Normalization of information sources (Eq. 3), 2) Integration of information sources (Eq. 4, where w=1/n), and 3) Feedback from the integration layer to the information sources (Eq. 5). An important difference between this Normalized Recurrence simulation and previous ones is that the model was not allowed to iterate until reaching a criterion, because duration of competition (e.g., reaction time) was not the measure of interest. Rather, the model was stopped at various intervals and the activations of the interpretation units were treated as probabilities of "grammatical" and "ungrammatical" responses. In order to prevent unnaturally high d-primes, each interpretation unit has a maximum of .95 activation in this first simulation.

With each iteration, the model gets more and more "confident" in one of these decisions. Of course, in the case of only two competing alternatives, the moment one decision is greater in activation than the other, it is obvious that (in this deterministic version of the competition algorithm) the *current* winner will be the *ultimate* winner. However, for simulating the time course of information integration, we need to allow the model to settle toward some criterion activation, especially if we consider the possibility that different response mechanisms (e.g., manual response, vocal response, or eye movements) may have different criteria for execution.

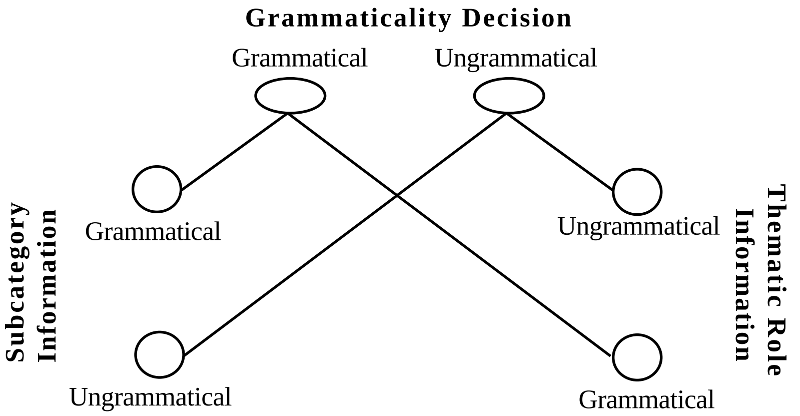

Figure 6. Schematic diagram of the Normalized Recurrence model designed to simulate the results of McElree and Griffith (1995).

In the first simulation of McElree and Griffith's (1995) SAT version of the speeded grammaticality judgment task, the model was given input values indicating either a grammatical sentence, subcategory violation sentence, or thematic violation sentence. The model was then allowed to iterate, gradually converging toward a decision on the grammaticality of the input, until an interruption point was reached, at which time the integration layer's values were recorded for the probability of a correct response (as though the model were being interrupted and forced to make a decision). For grammatical sentences, the input value for the grammatical node in each constraint was .51, and thus the input value for the ungrammatical node in each constraint was .49. When each iteration is treated as 50 ms, these values produce "grammatical" response times that approximate those from McElree and Griffith (1995). For a subcategory violation, the input values for the subcategory nodes were .2 grammatical and .8 ungrammatical, whereas for a thematic violation, the input values for the thematic role nodes were .4 grammatical and .6 ungrammatical.

To compute d-primes at each time step of the model, the activation of the "grammatical" integration node after a grammatical input was treated as the percentage of *hits*, and the activation of the "ungrammatical" integration node after ungrammatical input was treated as the percentage of *correct rejections*. Figure 7 compares McElree and Griffith's data to the model's d-prime values as a function of processing time. The first thing to notice is that the model reaches asymptote much more abruptly than in the human data. This is primarily due to a .95 maximum imposed on the activations in order to prevent d-primes of 4+. Much of the smooth, graded approach to asymptote exhibited by Normalized

Recurrence actually takes place between .95 and 1.0 activation. With that range omitted, this first simulation rather suddenly hits a sharp maximum before it is through with the steeply rising portion of its sigmoid function over time. Despite this obvious weakness of the first simulation, the critical portion of the data, where the early measurements for subcategory and thematic role violations are dissociated, is well accounted for by the model. Whereas McElree and Griffith's (1995) account of the data assumes that the curves for subcategory and thematic violations must depart from zero d-prime (or "become operative") at different points in time, Normalized Recurrence accounts for this portion of the data using two sigmoidal curves that "become operative" at the same time, but one has a stronger initial bias backing it up.

Improvements on this first simulation can be achieved in a number of ways. There are essentially six parameters in this model that can be manipulated: 1) the "grammatical input" value, 2) the "subcategory violation" value, 3) the "thematic violation" value, 4) the weights (since each pair must sum to 1, each of these first four terms counts as a single model parameter), 5) the activation rectification limit, and 6) the amount of time each iteration corresponds to. In the first simulation, the space of parameters 2 and 3 was searched (in steps of .05) to converge on an approximate fit to the data.

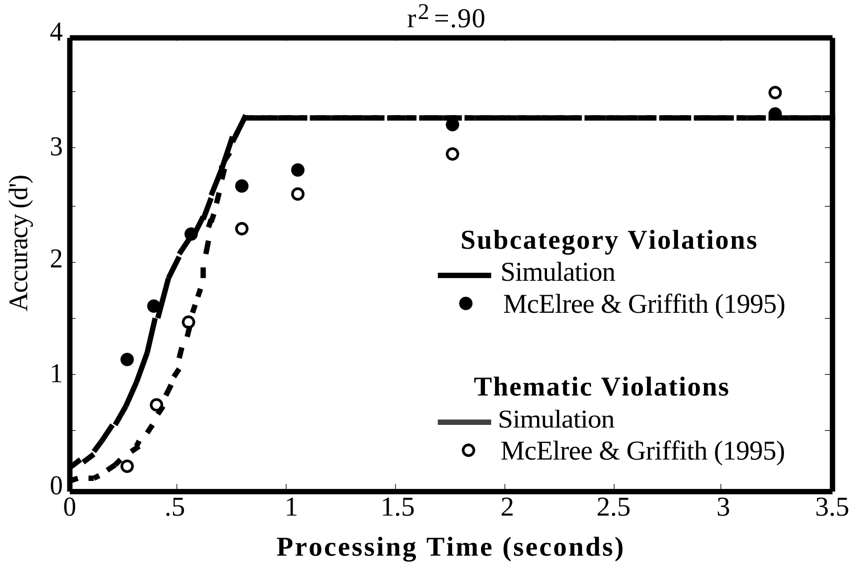

Figure 7. Accuracy of grammaticality detection (McElree & Griffith, 1995) and results of the first simulation with Normalized Recurrence. The fit to the data accounts for 90% of the variance.

In this next simulation (Figure 8), parameters 5 and 6 were modified to converge on a fit to the data. The input values for the different experimental conditions were identical to those of the first simulation. However, instead of a strict activation rectification, the normalization function (Eq. 3) added a small uniformly random value between 0 and .2 to the denominator at each time step (cf. Heeger, 1993). Also, the time constant was reduced to 30 ms per iteration.

The experimental results of McElree and Griffith's (1995) SAT version of the speeded grammaticality judgment task are certainly intriguing. Unlike most experimental methodologies in the field of sentence processing, the SAT procedure provides a window into preliminary incomplete representations that are in the process of being computed as information continuously accrues. However, attempting to extrapolate from the sampled d-primes to the underlying function's x-intercept via an exponential function may prematurely imply separate discrete points in time at which different linguistic processors "become operative." Instead, the results of these simulations suggest that the sampled d-primes over time may come from a system that integrates its different information sources simultaneously but with differing strengths. A weaker

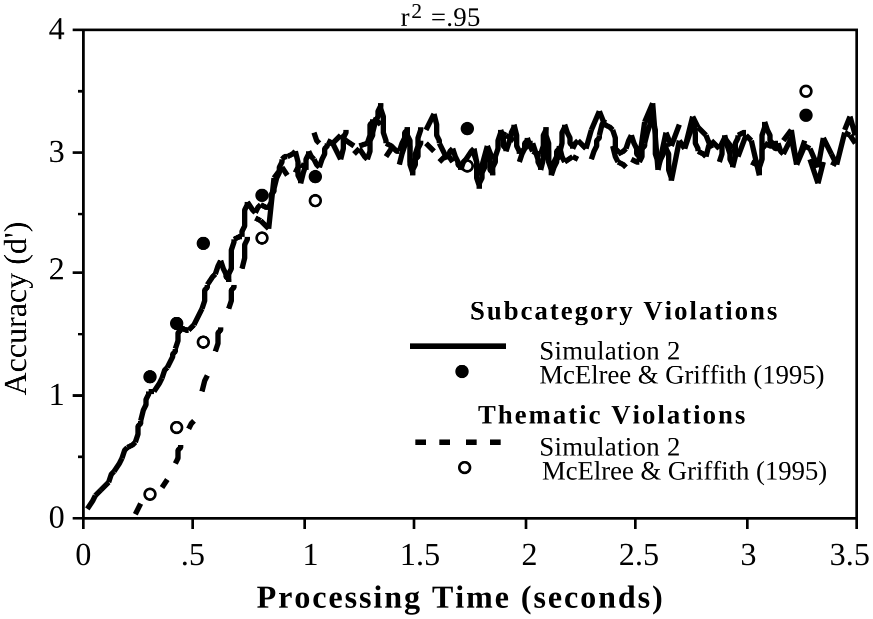

Figure 8. Accuracy of grammaticality detection (McElree & Griffith, 1995) and results of the revised simulation with Normalized Recurrence. The fit to the data accounts for 95% of the variance.

signal (e.g., thematic constraints) that "becomes operative" *at the same time* as a stronger signal (e.g., subcategory constraints) will still take longer to rise above the noise inherent in a probabilistic information processing system (Figure 8).

Although the performance of the Normalized Recurrence model is encouraging, the simulations presented here did not quite account for as much of the variance in the data as did McElree and Griffith's (1995) six-parameter Stepit-driven exponential fit (Eq. 5). Moreover, McElree and Griffith's (1998) more recent findings hold still more challenges for a parallel processing system, such as syntactic island constraints having higher d-primes than lexically-specific constraints, and crossings between different curves of d-prime over time. Future work with this model will explore further manipulation of the parameters of this network.

A Prediction from Competition

Many competition-based models (and other dynamical models) of sentence processing assume that a representation's activation will have a relatively non-extreme value during early moments of processing, and will gravitate toward an extreme value (e.g., minimum or maximum) as time proceeds -- modulo the occasional nonmonotonic vacillation. Note that, since its representations are localist nodes, Normalized Recurrence's attractors are corners in the state-space, and therefore a single run of the model with only two competing interpretations cannot exhibit vacillations. (Nonmonotonic behavior on one run of the model, such as that in Panel C of Figure 2, can only happen when several information sources compete over several interpretations, or when some stochasticity is added to the normalization function.)

When only two interpretations are competing in Normalized Recurrence, as one begins to increase in activation, the other must decrease, and they will continue on these trajectories monotonically. For example, with the ambiguity between a Main Clause (MC) and a Reduced Relative (RR) (3), input to the model that averages just barely in favor of the MC will cause the model to start out with equibiased representations for the MC and RR that gradually settle entirely in favor of the MC interpretation. In contrast, a model that posits a separate processing stage for syntactic biases followed by a stage for thematic role biases (e.g., Frazier & Fodor, 1978; Frazier, 1987; McElree & Griffith, 1998) might predict zero activation of the RR representation early on, regardless of what thematic fit information suggests. If thematic role information strongly biases an RR interpretation (such as a *prisoner* being a good Patient and a poor Agent of a *capturing* event), the activation of the RR representation will, at later points in time, eventually accrue some positive activation.

(3) a. The prisoner captured *a rat and kept it as a pet.* (Main Clause)
b. The prisoner captured *by the guards was tortured.* (Reduced Relative)

Thus, the prediction made by Normalized Recurrence, and ruled out by the two-stage models, is the following: With sentence fragments of the form "The"-noun-verb"-ed",in which thematic role information strongly biases the RR structure, even early moments of processing should show nonzero activation of the RR representation. A further, more specific, prediction from Normalized Recurrence is that those particular sentence fragments in which all constraints conspire *just barely* in favor of the MC, should in fact elicit greater positive activation of the RR representation during the *early* moments of processing than during the *later* moments of processing.

To test these predictions, we have designed a novel experimental methodology: speeded sentence completions, in which participants read sentence fragments, one word at a time, and complete these sentences under various time constraints. They are allowed 300 ms, 600 ms, 900 ms, or 1200 ms to prepare the completion. The results from 63 participants are presented herein.

Speeded Sentence Completions

Each trial proceeded as follows: a red circle appeared in the center of the screen indicating where the words would be presented, each word of the sentence fragment was presented in a noncumulative fashion in the center of the screen for 500 ms, three periods then appeared indicating that the participant should start preparing a completion, then a green circle appeared indicating that a completion must begin within 300 ms. Participants found the task difficult at first, but a 20-trial practice session (with a 500 ms processing interval) was typically enough to acquaint the participant with the task.

In this experiment, the three periods were on the screen for 300, 600, 900, or 1200 ms. These four different processing-interval conditions were run as separate blocks. In each block of 20 trials, the first 10 were fillers(ranging from two to four words in length), allowing the participant to get accustomed to that particular processing-interval condition. The remaining 10 trials in the block had four critical sentences embedded among 6 fillers items. The order of these blocks was randomized for each participant.

Participants were instructed to speak into the microphone what first came to mind and not to censor themselves. They heard a beep if they started responding too soon (i.e., while ‘...’ was still on the screen), and saw a “Respond faster!” sign on the screen if they began their response more than 300 ms after the green circle appeared. After finishing a sentence, they pressed a key on the button box to advance to the next trial.

The critical sentences were constructed from sixteen verbs, each with a typical Agent and a typical Patient for that particular event. Agenthood and Patienthood ratings were taken from norms collected in the work of McRae et al. (1998), and the verb form frequencies (Simple Past Tense, Past Participle, and Base frequency) were taken from Kucera & Francis (1982). See Table 1.

Table 1. Stimuli used in Speeded Sentence Completions

	Verb Frequency				Thematic Fit			Thematic Fit	
VERB	SPast	PPart	Base	NOUN1	Ahood	Phood	NOUN2	Ahood	Phood
arrested	4	15	27	police	6.45	1.46	suspect	1.40	5.49
audited	0	1	3	government	6.17	3.00	taxpayer	2.72	6.16
captured	2	15	33	troops	5.97	3.87	prisoner	1.76	5.03
convicted	1	13	16	juror	6.61	1.32	criminal	1.45	5.87
cured	1	6	20	doctor	6.76	3.78	patient	1.37	6.14
executed	1	13	22	terrorists	6.05	4.03	hostages	1.66	4.95
graded	0	2	3	teacher	6.94	2.60	student	2.42	6.81
instructed	2	14	23	coach	6.74	2.11	trainee	1.66	6.22
investigated	2	16	38	auditor	6.25	2.22	theft	1.22	6.78
paid	1	95	256	man	5.50	3.65	tax	1.63	5.43
punished	1	8	14	parent	6.50	1.54	child	1.53	5.78
rescued	1	5	14	knight	5.97	1.68	victim	1.21	4.89
sent	1	74	172	manager	5.55	2.95	package	1.58	6.16
sentenced	1	8	9	judge	6.94	1.27	defendant	1.25	6.35
tortured	1	8	10	kidnapper	5.68	1.60	slave	1.29	5.57
worshipped	1	2	12	priest	6.67	4.05	goddess	1.50	6.73

To utilize as many data as possible, all responses that began 300-1500 ms after the onset of the three periods were included in the analysis. Responses that were too early or too late for their condition were counted as belonging to the temporally accurate processing-interval. For example, if during a block of trials with the 600 ms delay a response occurred at 950 ms, it was counted as belonging to the 900-1200 ms processing-interval bin. Responses that began after 1500 ms (3%) and responses that were incomplete and/or still ambiguous (6%) were excluded from analysis.

The overall results of this study are compelling. At the earliest measured point in time, the 300-600 ms bin, sentence fragments with Patient-like nouns show significantly more reduced relative completions than those with Agent-like nouns (25% vs. 2%; p<.05). Figure 9 shows the percentage of RR completions for both good Patients and good Agents at the four processing-interval bins. When the sixteen items are averaged for each curve, the temporal dynamics from individual items cancel each other out, resulting in relatively flat curves that are consistently about 25% apart from one another. We do not see in the good Patient condition an initial near-zero percentage of RRs that gradually increases over time, as would be most naturally predicted by a syntax-first model. Nonetheless, although this result seems most consistent with a simultaneous integration of constraints account of sentence processing, a syntax-first model can always accommodate these findings by restricting the purely syntactic processing stage to the first 300 ms of processing.

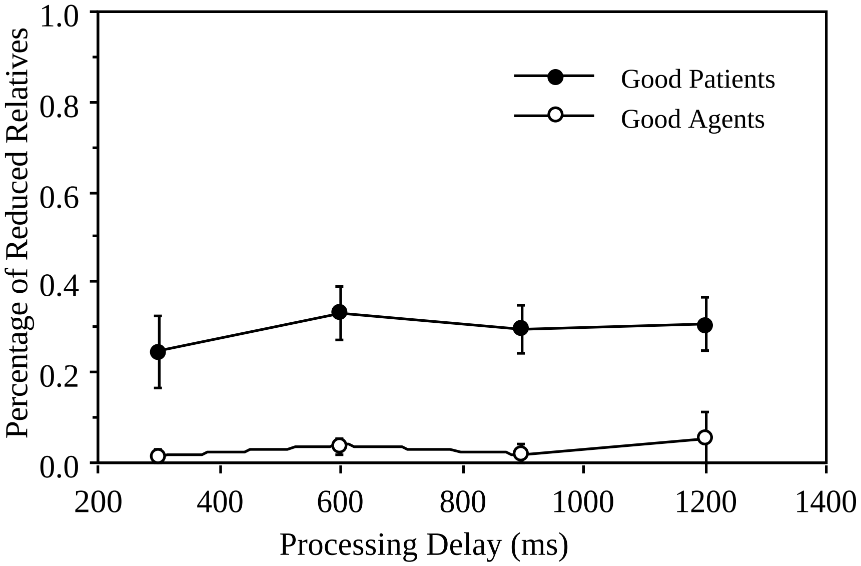

Figure 9. Overall results of the speeded sentence completion task.

The more specific prediction made by competition models is also borne out: that a particular sentence fragment in which all constraints conspire *just barely* in favor of the MC will elicit greater positive activation of the RR representation during the *early* moments of processing than during the *later* moments of processing. For example, "The prisoner captured..." elicited 35-40% RRs during the early delay conditions, and 0-10% RRs during the latter delay conditions. Syntax-first models are fundamentally incapable of explaining such a result, whereas Normalized Recurrence predicts this result quite naturally.

Normalized Recurrence

As the Normalized Recurrence competition algorithm emerged in the context of the constraint-based lexicalist framework in sentence processing (e.g., Filip et al., this volume; McRae et al., 1998; Spivey & Tanenhaus, 1998), it makes sense to apply the model to the lexically specific stimuli used in this experiment and average the two groups of 16 runs of the model for comparison with the averaged human data (Figure 9). (In fact, very different and inappropriate results would arise from instead averaging the stimulus parameters in the two groups of 16 items and running the model twice with those averaged values.)

Figure 10 shows a schematic diagram of the Normalized Recurrence simulation of the speeded sentence completions. We used the same three

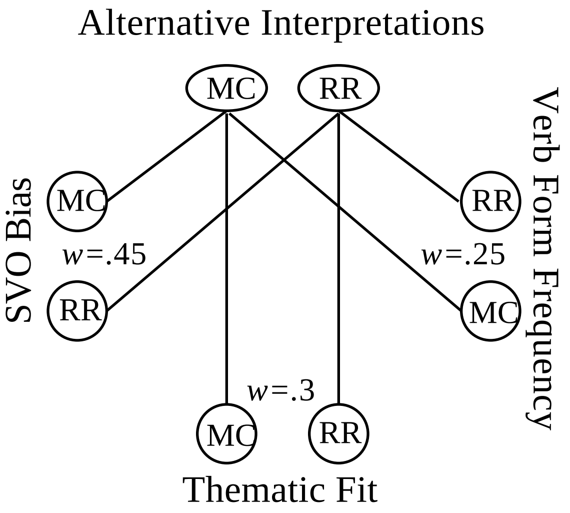

Figure 10. A schematic diagram of the Normalized Recurrence model that simulates the speeded sentence completions data.

information sources as in McRae et al. (1998): SVO bias, Thematic Fit, and Verb Form Frequency. For the SVO bias, MC=.92 and RR=.08 (McRae et al., 1998). For the lexically specific biases, the values were taken from Table 1. Thematic fit ratings were entered "as is," and the verb form frequencies were entered as MC=SPast/Base, RR=PPart/Base. (For the two verbs where SPast=0, the values were entered as MC=.01 and RR=.99, instead of MC=0 and RR=1.) After searching the weight-space for this network (in steps of .05), the best approximate fit to the data was found with the SVO Bias being weighted at .45, Thematic Fit weighted at .3, and Verb Form Frequency weighted at .25. Although the weight for Verb Form Frequency is notably greater here than in McRae et al. (1998), the ordinal ranking of McRae et al's weights is preserved.

The model was given input from all 32 noun-verb pairs, and allowed to iterate for 120 cycles of competition, treating each iteration as equivalent to 10 ms of processing time. Thus, the activation of the RR Interpretation node from cycle 30 to 120 provided the model's prediction of the probability of an RR completion during the four processing-interval bins in Figure 9.

When the model's results from the 16 good Patient items were averaged, they slightly overestimated the percentage of RR completions, at around 40%. See Figure 11. Similarly, at the first processing-interval bin, the model slightly overestimated the percentage of RR completions for good Agent items as well. Notably, however, just as the temporal dynamics of the individual items canceled

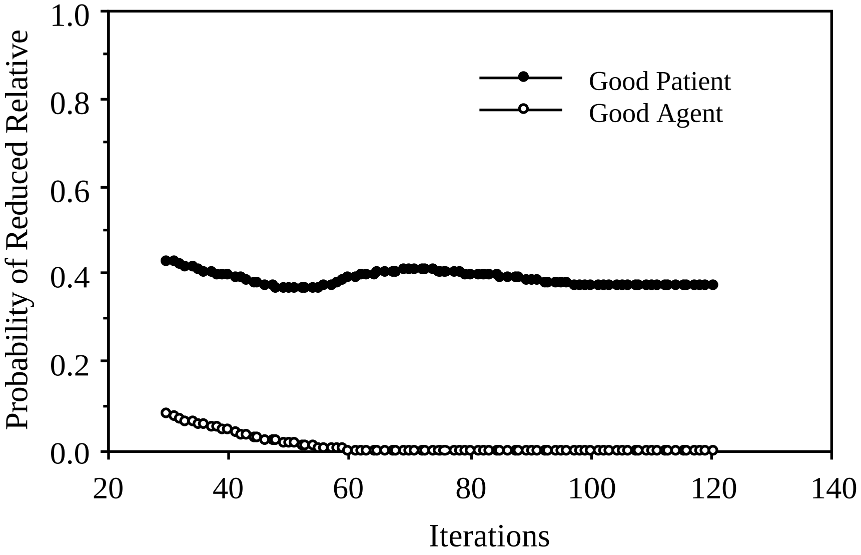

Figure 11. Results of Normalized Recurrence's simulation of the speeded sentence completion task, averaged across the 16 verbs.

each other out when averaged in the human data, so did the temporal dynamics of the individual items in the model simulations cancel each other out when averaged. The model's fit to the human data when averaged across participants and items is close: r^2=.92.

Future work will need to break down these two curves into their item-by-item effects, and test the model's account of the behavior of individual sentence fragments. Since the constraint-based lexicalist framework predicts systematic item-by-item variation, the ultimate challenge for this account of sentence processing is to simulate the temporal dynamics of individual stimulus items. As the typical dataset in a sentence processing experiment contains perhaps 4-5 data points per stimulus item per condition, this new goal will require a much larger than usual dataset.

Until now, serial stage accounts of sentence processing have enjoyed the position of needing only to demonstrate effects averaged across items. However, as these theories become more explicit in their account of how the later stages work, they too will need to make predictions about item-by-item variation, e.g., handled by a late constraint-based stage or by a rule-based reanalysis system.

General Discussion

In this chapter, we have discussed the benefits of a few new tools in sentence processing, both theoretical and methodological. Nonlinear dynamics provides a new perspective for understanding the simultaneous existence of systematic, rule-like behavior in language, via nearby strong attractors, and sporadic, probabilistic behavior in language, via distant or weak attractors (cf. Tabor & Hutchins, 2000). Most dynamical models of sentence processing generally posit that all available constraints on interpretation are active simultaneously, but with varying strengths -- and the results of these strength differences, as the system gravitates toward an attractor, can be quite nonlinear. Clearly, the best way to test this kind of account of language is to explore the temporal dynamics of language processing at a fine-grain scale, and look for the kinds of nonlinearities that are predicted.

In contrast to dynamical models, serial stage models of sentence processing tend to account for rule-like constraints and more probabilistic constraints with completely separate processing systems that apply their constraints at different points in time (Frazier & Fodor, 1978; Frazier, 1987; McElree & Griffith, 1995, 1998). In support of this kind of account, the results of McElree and Griffith's (1995, 1998) SAT procedure with the speeded grammaticality judgment task show what look like differential "start times" for syntactic processing, verb-subcategory processing, thematic role processing, etc. However, simulations with the Normalized Recurrence competition algorithm demonstrate that McElree and Griffith's functions of d' over time can be approximated by a model that integrates all information sources simultaneously, just with different input strengths. Essentially, this amounts to an existence proof, showing that data that might have been interpreted as consistent only with a serial stage account of sentence processing may in fact be accommodated by a parallel, integrative dynamical model of information integration.

The next step comes when this "existence proof" makes a specific prediction: that at a point of syntactic ambiguity, *early* moments of processing will show partial activation of the non-preferred alternative -- and in some circumstances may even show greater activation of that alternative during early moments of processing than during later moments of processing. In order to test this prediction, a new methodology was introduced. Participants were instructed to complete sentence fragments (that were ambiguous between beginning a main clause or reduced relative clause) under varying time pressure. Results indicated that when semantic information supported the reduced relative, participants exhibited a substantial salience of the reduced relative alternative even at the earliest measured point in time. Moreover, with sentence fragments for which the constraints just barely favored the main clause, a reduced relative completion was more likely early on than later on. A simulation of Normalized Recurrence approximated these results rather well.

In sum, the evidence for serial stage models of sentence processing is waning. Many of the findings that were once treated as evidence that the influence of semantic information on parsing is delayed are being accommodated by models that apply syntactic and semantic biases simultaneously (e.g., Filip et al., this volume; McRae et al., 1998; Spivey & Tanenhaus, 1998; Tabor et al., 1997; Tanenhaus et al., 2000). Moreover, we report here suggestive evidence in the salience of syntactic alternatives for a type of temporal dynamics -- early activation of the non-preferred alternative which then decreases over time -- that is typically ruled out by serial stage models of sentence processing.

The goal here is not (not yet, anyway) to make it impossible to delineate what information sources are fundamental to sentence processing and what information sources are better treated as belonging to "the rest of perception and cognition." It is relatively clear that syntax is "fundamental", verb-subcategory information is "crucial", thematic role information is "pretty important", etc. As vague as those descriptors sound in distinguishing the relative import of each information source for sentence processing, so perhaps should the distinctions between the importance of these information sources in our models of sentence processing be vague. Instead of seeking evidence for discrete, qualitative architectural differences between these information sources, such as differential "start times," we advocate seeking quantitative strength differences between them, such as graded constraint weights, and a generic integration algorithm that they follow.

Acknowledgments

Supported by a Sloan Fellowship in Neuroscience to the first author. The authors are grateful to Mike Tanenhaus, Brian McElree, Daniel Richardson and one anonymous reviewer for helpful discussions, Melinda Tyler for assistance with data collection, and Jim Magnuson for help with the verb frequency data.

References

Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419-439.

Chandler, J. (1969). Subroutine STEPIT -- finds local minimum of smooth function of several parameters. *Behavioral Science, 14*, 81-82.

Clifton, C., Frazier, L., & Rayner (1994) (Eds.) *Perspectives on sentence processing*. Hillsdale, NJ: Erlbaum.

Cottrell, G. & Small, S. (1983). A connectionist scheme for modeling word sense disambiguation. *Cognition and Brain Theory, 6*, 89-120.

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195-225.

Ferreira, F. & Clifton, C. (1986). The independence of syntactic processing. Journal of Memory and Language, 25, 348-368.

Filip, H., Tanenhaus, M., Carlson, G., Allopenna, P., & Blatt, J. (this volume). Reduced relatives judged hard require constraint-based analyses.

Fodor, J. A., Bever, T., & Garret, M. (1974). *The psychology of language*. New York: McGraw-Hill.

Francis, W. & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Frazier, L. & Fodor, J. D. (1978). The sausage machine: A two-stage parsing model. *Cognition*, *6*, 291-325.

Frazier, L. (1987). Theories of syntactic processing. In J. Garfield (Ed.), *Modularity in knowledge representation* . Cambridge, MIT Press.

Green, D. & Swets, J. (1966). *Signal detection theory and psychophysics* . NY: Wiley.

Heeger, D. (1993). Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. *Journal of Neurophysiology, 70*, 1885-1898.

Henderson, J. (1994). Connectionist syntactic parsing using temporal variable binding. *Journal of Psycholinguistic Research, 23*, 353-379.

Kawamoto, A. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, *32*, 474-516.

MacDonald, M. Pearlmutter, N & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101,* 676-703.

McClelland, J. & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

McClelland, J. & Kawamoto, A. (1986). Mechanisms of sentence processing: Assigning roles to constituents of sentences. In McClelland & Rumelhart (Eds.), *Parallel Distributed Processing, vol. 2*. Cambridge: MIT Press.

McClelland, J. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287-330.

McElree, B. & Griffith, T. (1995). Syntactic and thematic processing in sentence comprehension: Evidence for a temporal dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 134-157.

McElree, B. & Griffith, T. (1998). Structural and lexical constraints on filling gaps during sentence comprehension: A time-course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 432-460.

McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1998). Modeling the effects of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 37,* 283-312.

Selman, B. and Hirst, G. (1985). A rule-based connectionist parsing system. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Spivey, M. & Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1521-1543.

Spivey-Knowlton, M. & Allopenna, P. (1997). *A computational account of the integration of linguistic and visual information in spoken word recognition.* Paper presented at the Computational Psycholinguistics Conference.

Spivey-Knowlton, M. & Sedivy, J. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition, 55,* 227-267

St. John, M. & McClelland, J. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 217-257.

Stevenson, S. (1994). Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research, 23*, 295-322.

Swinney, D. (1979). Lexical access during sentence comprehension *Journal of Verbal Learning and Verbal Behavior, 18,* 645-659.

Tabor, W. & Hutchins, S. (2000). Mapping the syntax/semantics coastline. In L. Gleitman & A. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. (pp. 511-516). Erlbaum: Mahwah, NJ.

Tabor, W., Juliano, C. & Tanenhaus, M (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes, 12,* 211-271.

Tanenhaus, Leiman, J. & Seidenberg, M. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior, 18*, 427-440.

Tanenhaus, M., Spivey-Knowlton, M., & Hanna, J. (2000). Modeling the effects of discourse and thematic fit in syntactic ambiguity resolution. In M. Crocker, M. Pickering, & C. Clifton (Eds.) *Architectures and Mechanisms for Language Processing.* (pp.90-118). Cambridge: Cambridge U. Press.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K. & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science, 268,* 1632-1634.

Trueswell, J. & Tanenhaus, M. (1994). Toward a lexicalist approach to syntactic ambiguity resolution. In C. Clifton, L. Frazier & K. Rayner (Eds.), *Perspectives on sentence processing*. Hillsdale, NJ: Erlbaum.

Waltz, D. & Pollak, J. (1985). Massively parallel parsing: A strongly interactive model of language interpretation. *Cognitive Science, 9*, 51-74.

Wiles, J. & Elman, J. (1995). Learning to count without a counter: A case study of dynamics in recurrent networks. *Proceedings of the 17th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.