LANGUAGE AND COGNITIVE PROCESSES, 1997, *12* (2/3), 211–271

Parsing in a Dynamical System: An Attractor-based Account of the Interaction of Lexical and Structural Constraints in Sentence Processing

Whitney Tabor, Cornell Juliano and Michael K. Tanenhaus

University of Rochester, Rochester, NY, USA

A dynamical systems approach to parsing is proposed in which syntactic hypotheses are associated with attractors in a metric space. These attractors have many of the properties of traditional syntactic categories, while at the same time encoding context-dependent, lexically specific distinctions. Hypotheses motivated by the dynamical system theory were tested in four reading time experiments examining the interaction of simple lexical frequencies, frequencies that are contingent on an environment defined by syntactic categories, and frequencies contingent on verb argument structure. The experiments documented a variety of contingent frequency effects that cut across traditional linguistic grains, each of which was predicted by the dynamical systems model. These effects were simulated in an implementation of the theory, employing a recurrent network trained from a corpus to construct metric representations and an algorithm implementing a gravitational dynamical system to model reading time as time to gravitate to an attractor.

INTRODUCTION

Bever's pioneering work (e.g. Bever, 1970) established that readers and listeners have strong preferences for certain syntactic sequences. For example, a noun phrase–verb–noun phrase sequence at the beginning of a sentence is typically taken to be a main clause. These preferences are revealed in garden-path effects for sentences with temporary syntactic

Requests for reprints should be addressed to Michael K. Tanenhaus, Department of Brain and Cognitive Sciences, Meliora Hall, University of Rochester, Rochester, NY 14627, USA. E-mail: mtan@bcs.rochester.edu

This research was supported in part by NIH Grant HD-27206 and by NIDCD Grant T32DC0035. Robbie Jacobs provided helpful feedback at various stages of this project. The final manuscript benefited from insightful comments by Don Mitchell, Dave Plaut and an anonymous reviewer. We did not include appendices with the materials used in Experiments 1–4 because of length constraints. These are available on request from M.K.T.

amgibuities that do not conform to these preferred syntactic configurations, as illustrated in the examples in (1):

1a. The horse raced past the barn fell.
1b. The woman warned the lawyer was misguided.

Structural preferences like these inspired a family of parsing models in which category-level structure guides the initial syntactic structuring of the linguistic input (for a review, see Tanenhaus & Trueswell, 1995). For example, in an influential proposal of this class, the two-stage, or "garden path" model, developed by Frazier and colleagues (Frazier & Rayner, 1982; Frazier, 1987), initial structure is assigned using only syntactic category information, and a small set of structurally defined parsing principles such as minimal attachment or late closure. Other information is used to evaluate, filter and, if necessary, revise the initial structure. Information excluded from initial consideration includes lexically specific syntactic and semantic information (e.g. information about the types of complements licensed by a verb and the semantic properties or thematic roles associated with these complements).[1] Evidence in support of these claims comes from numerous studies demonstrating increased processing difficulty in the form of elevated reading times when the structure that is ultimately correct is consistent with lexical constraints but inconsistent with category-based parsing principles (for reviews, see Frazier, 1987; Mitchell, 1989; Mitchell, Cuetos, Corley, & Brysbaert, 1995). Thus elevated reading times reflect syntactic misanalysis and the time it takes the reader to recover from the misanalysis.

More recently, however, a body of research has emerged demonstrating that lexically specific syntactic and semantic information, including the frequencies with which lexical items occur in different environments, can reduce, and sometimes eliminate, many strong structural preferences (for recent reviews, see MacDonald, Pearlmutter, & Seidenberg, 1994; Tanenhaus & Trueswell, 1995). For example, the sentences in (2) illustrate that, for both of the types of ambiguities in (1), there are sentences with the same structure, but with different lexical items, that are read with little or no processing difficulty (Garnsey et al., 1995; Trueswell, 1996; Trueswell, Tanenhaus, & Garnsey, 1994; Trueswell, Tanenhaus, & Kello, 1993).

2a. The land mine buried in the sand exploded.
2b. The woman said the idea was misguided.

Several classes of theoretical alternatives can be adopted to account for these lexical influences. One alternative is to retain a two-stage architecture

[1] Other two-stage models allow lexically based syntactic constraints to affect first-stage parsing, but still exclude non-syntactic lexical constraints and information such as frequency that is not incorporated into traditional grammars (e.g. Pritchett, 1992).

to capture clear cases of structural infuences. Lexical influences are then attributed to rapid revision effects (Frazier, 1989; Mitchell, 1989), with processing difficulty reflecting the time it takes to complete the revision. However, two-stage models do not have a principled way of accounting for why lexical influences are strong enough to completely mask putative first-stage effects in some syntactic environments but not others.

A second, more promising alternative is to treat structural and lexical influences as arising from a system that integrates constraints from independent levels of representation; for example, structural biases are couched in terms of sequences of categories (phrase structure rules) and lexical biases, including preferences for syntactic categories, argument structures, thematic structures, and so forth. A sensible first hypothesis is to assume that this information is combined in a probabilistic model. At each point in the sentence, the conditional probabilities of the possible continuations are computed. Processing load is then inversely related to the probability of occurrence of the actual sequence encountered; that is, low-probability sequences are relatively difficult to process.[2] Conditional probability models provide a theoretical basis for incorporating probabilistic lexical information into a model that uses syntactic rules. However, they do not provide insight into the systematic variation across contexts of the relative strengths of structural and lexical constraints. Standard conditional probability models also have a hybrid architecture in that they assume representations that are intrinsically discrete (e.g. rules in a symbolic system), and then impose probabilities on these rules. This begs important questions about how the representations in the system emerge during learning. These issues are closely related to the questions about how to define and combine information at different linguistic grains (cf. Mitchell et al., 1995).

This paper explores a third alternative, which we believe has the potential to provide revealing insights into both the nature of linguistic representations and the processing dynamics that operate on these representations. The basic idea following similar proposals by Elman (1993), Kawamoto (1993) and others is inspired by dynamical systems theory and representational ideas drawn from the connectionist learning literature. We assume that processing involves following a trajectory through a metric space, structured according to a similarity principle, such that words that are likely to lead to similar continuations are close to one another in representational space. Categorically distinct states of the language processor are associated with distinct *attractors*. Processing involves using the information provided

[2] Charniak (1993) reviews automatic parsing models along these lines; Jurafsky (1996) describes a conditional probability model which assigns preferences to ambiguous sequences that are broadly consistent with experimental results in the psycholinguistic literature.

by each successive word to place the processor somewhere in an attractor space and then letting it gravitate to whatever attractor manages to capture it. Successful parsing of each partial word sequence corresponds to arriving at, or getting very near, a single attractor (so there is no uncertainty about the current categorical status). The path followed by the trajectory through the attractors represents the syntactic structure of the sentence. Processing time is modelled as time taken to gravitate to such a state.

While this system can, in principle, be applied to both syntactic category ambiguities and attachment ambiguities, involving phrase-level structure, this paper will focus on syntactic category ambiguities. The hypothesis we are exploring is that category-based parsing preferences are embedded in representations that *emerge* within a constraint-based learning system because of similarities among lexical items in particular syntactic contexts (e.g. Juliano & Tanenhaus, 1994). The hypothesis is that an appropriate balance between lexcial and structural influences will arise as a result of (a) the nature of the lexical input, (b) the learning system that creates representations based on this input, and (c) the processing dynamics of the system.

The remainder of the paper is organised into six sections. First, we provide a brief introduction to some important constructs of dynamical systems theory, emphasising the theoretical constructs we apply to syntactic processing. The next section then presents three self-paced reading experiments that examine the interaction of contingent (i.e. conditional) frequencies at different linguistic grains to test some hypotheses derived from our approach. These experiments also serve as the empirical base for evaluating an implementation of our theory, which is presented in the following section. We implement the model with a dynamical system that uses representations developed by a recurrent neural network trained on word prediction (Elman, 1990, 1991) with inputs from small finite-state grammars. The penultimate section presents a model trained from a representative sample of a real corpus and uses it to model verb-specific variance in reading times for an experiment also reported in this section. The final section summarises our results and discusses the strengths and limitations of our models, as well as alternative approaches.

PARSING IN A DYNAMICAL SYSTEM

Dynamical systems theory is typically concerned with systems that change continuously with time (for good introductions, see Abraham & Shaw, 1984; Perko, 1991; Strogatz, 1994). Examples include: swinging pendulums; orbiting stars and planets; populations fluctuating in an ecosystem; gases swirling around in an atmosphere. It is useful to consider the trajectories of a dynamical system—that is, the paths it can follow as time progresses. In the case of the rigid-arm pendulum, some trajectories swing back and forth, others whirl around the circle, and two of them remain at one point indefinitely (hanging down and, improbably, balanced straight up). If the pendulum is damped, then all trajectories except those leading to the improbable state approach the low point-trajectory in the limit. Such a limiting trajectory is called an *attractor*. Those starting points from which the system gravitates towards a particular attractor A are collectively referred to as the *basin of attraction* of A. The basin of attraction of the pendulum's low attractor consists of every state except those that lead to the improbable state. In the case of a planetary system, each large mass is surrounded by its own basin of attraction. If a dynamical system is near one attractor and far from others, its behaviour is dominated by that attractor. By contrast, if the system is in a position intermediate between several attractors, it shows all of their influences simultaneously. We will refer to this property of attractors as the *local dominance* property. Local dominance is easy to understand intuitively in the case of a planetary system: A satellite near one large mass behaves, for all practical purposes, as though that mass is the only mass around, but a satellite roughly equidistant from several masses moves on a complex trajectory which reflects their simultaneous influences. We take advantage of the local dominance property of attractors to model the interaction of syntactic and lexical influences on sentence processing.

We hypothesise that categorically distinct states of the language processor are associated with distinct attractors in a metric space that is structured according to a similarity principle. Partial word sequences occupy nearby positions in the space if they are likely to have similar continuations. Processing involves using the information provided by each successive word to place the processor somewhere in an attractor space and then letting it gravitate to whatever attractor manages to capture it. Successful parsing of each partial word sequence corresponds to arriving at, or getting very near, a single attractor (so there is no uncertainty about the current categorical status). Processing time is modelled as the time taken to gravitate to such a state. The syntactic structure of the sentence can be represented by the succession of attractors that the processor visits during the processing of the sentence.

This similarity-based processor is closely related to a symbolic processor based on a grammar of the language. In particular, corresponding to each state, S[i], of the symbolic processor, there is a continuous (connected) region, R[i], in the metric space. Importantly, the use of regions instead of states allows the processor to be sensitive to quantitative distributional differences among elements belonging to the same lexical class. These differences give rise to small, within-region contrasts. For our current purposes, the most important property of the metric space representation is that syntactic category ambiguity is associated with *representational* *intermediacy*, as in Kawamoto's (1993) model of lexical ambiguity resolution. The following two examples help illustrate this point:

Example 1. Temporarily ambiguous word sequences give rise to intermediate representations. For instance, the sentence fragment *the insect examined* ... is ambiguous between a past-participle and a main verb interpretation of the verb *examined*. Thus the fragment's representation in the metric space is intermediate between, for example, the representation of *the insect knew* ... and the representation of *the insect known* ... If, as seems likely in this case, nouns like *insect* more often function as the object of verbs like *examined* in a corpus used to derive the representation, then the fragment *the insect examined* ... will be placed closer to the representation of *the insect known* ... than to the representation of *the insect knew* ... A noun with the opposite distributional tendency (e.g. *entomologist*) will have the opposite representational leaning.

Example 2. Grammatically unambiguous cases involving a mixture of signals can give rise to representational intermediacy. For example, *that person* ... will be placed in the metric space between *this person* ... and *that people* ... (as in *That people starve is unacceptable*) but closer to *this person* ... However, the possible continuations are consistent with only the Determiner + Singular Noun pattern (*this person* ...) and so the representational leaning in this direction is extremely strong, even though the presence of the word *that* in *that person* produces a bias in the direction of the sentential subject interpretation. This bias will influence the representation in a small way after the disambiguating word *person* has been encountered.

We can use the attractor mechanism to achieve an appropriate balance between structural and lexical influences in a variety of syntactic contexts. We hypothesise that in contexts in which lexical influences are robust, the processor lands in a range of intermediate positions between attractors, and thus takes different amounts of time to reach certainty, depending on which lexical items are involved. In such cases, lexical manipulations alone are expected to determine the state of the processor. Thus the model has a way of handling the sensitivity to lexical differences which two-stage models have trouble with. Moreover, in contexts where lexical effects are more subtle, we expect the processor to land close to a single attractor every time the context is encountered. In implementations of the model, we use a connectionist learning network trained from a corpus to generate processing locations and an explicit gravitational algorithm to measure gravitation time to an attractor.

The dynamical framework leads to a uniform treatment of frequency effects in parsing that is similar in spirit to recent attractor-based accounts of frequency and consistency effects in word recognition (e.g. Plaut, McClelland, Seidenberg, & Patterson, 1996), and of syntactic category shifts during language change (Tabor, 1994, 1995). Superficially contradictory effects of linguistic generalisations at different levels of traditional linguistic structure or *grains* (Mitchell et al., 1995) emerge as natural properties of the model. Moreover, many local processing difficulties that are typically attributed to syntactic misanalysis receive an alternative interpretation as *competition among attractors*, even when the processing system adopts the correct analysis. This competition account is consistent with claims made by many constraint-based models (e.g. Bates & MacWhinney, 1989; Hanna, Spivey-Knowlton, & Tanenhaus, 1996; MacDonald et al., 1994; Spivey-Knowlton, Trueswell & Tanenhaus, 1993; Spivey-Knowlton & Sedivy, 1995; Spivey-Knowlton & Tanenhaus, submitted; Tanenhaus & Trueswell, 1995).

CONTINGENT FREQUENCIES AND *THAT*

To evaluate the hypotheses introduced in the previous section, we explored a linguistic domain which seemed promising for observing emergent category effects and lexical-category interactions. The word *that* has several propeties, making it a useful test case. *That* is ambiguous among multiple syntactic categories. For example, *that* can be a demonstrative determiner as in (3a), a complementiser as in (3b), a relative pronoun as in (3c), and a pronoun as in (3d). These category ambiguities can give rise to phrasal ambiguities, including the ambiguities in several structures that have figured prominently in the psycholinguistics literature.

3a. That experienced diplomat would be helpful to the lawyer.
3b. That experienced diplomats would be helpful made the lawyer confident.
3c. The experienced diplomat that the lawyer admired was helpful.
3d. The lawyer didn’t believe the experienced diplomat said that.

We will focus exclusively on the ambiguity between *that* as a complementiser and *that* as a demonstrative determiner, an instance of a lexical category ambiguity that leads to phrasal ambiguity. When used as a demonstrative determiner, *that* introduces a noun phrase; when used as a complementiser, *that* introduces a sentence complement. This is illustrated in examples (3a) and (3b). In sentence (3a), *that experienced diplomat* is the subject noun phrase of a main clause, whereas in (3b), *that experienced diplomats* is the beginning of an extraposed sentence complement. The number of the noun disambiguates *that* as either a complementiser or a determiner. With a plural noun, *that* must be a complementiser because *that* is singular and demonstrative determiners have to agree in number with the noun they specify. With a singular noun, *that* must be a demonstrative because the head noun in a bare noun phrase (i.e. a noun phrase without a specifier) must be plural.

The relative frequency with which *that* is used as a complementiser and a determiner varies with syntactic environment. According to the Francis and Kučera (1982) counts based on the Brown corpus, *that* is used more frequently as a complementiser (70%) than as a demonstrative determiner (15%). However our analysis of the Brown corpus revealed that, at the beginning of a sentence, *that* is more often used as a determiner (35%) than as a complementiser (11%), whereas after a verb, *that* is more often used as a complementiser (93%) than as a determiner (6%). Note that the statistics associated with *that* run counter to the more general bias for a determiner that follows a verb to introduce an NP-object complement. In addition, when *that* follows a verb, an environment in which it is typically a complementiser, the subcategorisation properties of the verb determine whether a complementiser analysis is grammatical. *That* can only be used as a complementiser when it follows a verb permitting a sentence complement. This is illustrated by the examples in (4) using the verb *insisted*, which permits a sentence complement, and *visited*, which does not:

4a. Bill insisted that experienced lawyer would be helpful.
4b. Bill insisted that experienced lawyers would be helpful.
4c. Bill visited that experienced lawyer.
4d. *Bill visited that experienced lawyers.

Predictions from the Attractor Framework

In our dynamical systems model, different syntactic environments correspond to different regions of the representation space, and the local attractor configuration in each region reflects the context-dependent frequencies of occurrence of the different categorisations. Therefore, a reader's bias to interpret *that* as a complementiser versus a determiner should be influenced by contingent frequencies rather than just by simple lexical frequencies. In particular, when *that* is encountered at the beginning of a sentence, it will be affected by two attractors corresponding to the complementiser and determiner classifications. *That* will be placed closer to the *that*-(Det)-N attractor because "determiner" is the most common category at the beginning of a sentence, and *that* occurs more often as a determiner in this position. This should be reflected in a strong bias to process *that* as a determiner, a prediction that was tested in Experiment 1. *That* at the beginning of a sentence, where the determiner reading is more common, was compared with *that* after a verb, where the complementiser reading is more common.

A second prediction follows from the observation that the word *that* is nearly always a complementiser when it follows a verb—processing should be influenced by a strong *that-as-complementiser* attractor in this environment. This attractor should be strong enough to influence processing even when *that* follows a verb that does not permit a sentence complement. Experiment 2 tested this prediction by contrasting verbs like *insisted*, which strongly prefer to take a sentence complement (SC-bias verbs), with verbs like *visited*, which do not permit a sentence complement and typically occur with an NP complement (NP-bias verbs).

Attractor effects that are analogous to those seen for lexical category ambiguities should also be observed for phrasal ambiguities. *The*, and other determiners, typically introduce an NP-object complement when they follow a verb. This should result in a strong *NP-Det-object* attractor. Verbs that typically take sentence complements should form an *Sbar-complement* attractor. However, the NP-object attractor should still influence the processing of *the* when it follows an SC-bias verb. This prediction was tested in Experiment 3.

Experiment 1: Contingent Frequency Versus Simple Lexical Frequency

This experiment examined the ambiguity between *that* as a complementiser (e.g. The lawyer said *that* cheap hotels would be safe) and *that* as a demonstrative determiner (e.g. The lawyer said *that* cheap hotel would be safe). Reading times to *that*-adjective-noun sequences when they occurred at the beginning of a sentence were compared to when they followed an SC-bias verb. The number on the noun disambiguated *that* as a complementiser or a determiner. A plural noun disambiguates *that* as a complementiser because demonstrative determiners must agree in number with the noun they specify, and *that* is a singular demonstrative. A singular noun disambiguates *that* as a demonstrative because the head noun in a noun phrase without a specifier (i.e. a bare noun phrase) must be plural. Sample materials are presented in (5):

5a. The lawyer insisted that cheap hotel was clean and comfortable.
5b. The lawyer insisted that cheap hotels were clean and comfortable.
5c. That cheap hotel was clean and comfortable to our surprise.
5d. That cheap hotels were clean and comfortable surprised us.

Reading times to the noun and the words immediately following should be shorter when the number of the noun is congruent with the syntactic category that the reader has assigned to *that*. When *that* is interpreted as a complementiser, reading times should be shorter after a plural noun compared to after a singular noun. The opposite pattern should be observed when *that* is interpreted as a demonstrative. Crucially, the preferred resolution of *that* should interact with syntactic environment. In sentence-initial position, where *that* is more typically used as a determiner, reading times should be shorter after a singular noun compared to a plural noun. After a verb, where *that* is more typically used as a complementiser, reading times should be shorter following a plural noun compared to a singular noun.

Method

Participants. Thirty-six students from the University of Rochester participated for course credit. All were native speakers of English.

Materials. Four sentences were constructed for each of 20 *that* + adjective + noun sequences, by varying two properties of the sentences: (1) the noun was either singular or plural, disambiguating *that* as a demonstrative or as a complementiser, respectively, and (2) the *that*-phrase either came at the beginning of the sentence or it followed a verb. In this experiment, the test sentences used only verbs that typically occur with sentence complements, and rarely, if ever, permit a noun phrase complement. Verb-type is varied in subsequent experiments.

Each trial consisted of two sentences, with the second sentence being a natural continuation of the first. On critical trials, the target sentence was always the first sentence in the trial. Four lists were constructed by assigning each of the four sentences created from a particular *that* + adjective + noun sequence to a different list. The target sentences were pseudo-randomly combined with 42 distractor trials, for a total of 62 trials in each list. At least one filler trial appeared before each experimental trial, and conditions were balanced across the two halves of the stimulus lists.

Procedure. Stimuli were presented one word at a time using a moving window presentation format (Just, Carpenter, & Woolley, 1982) on a PC computer equipped with a Digitry CTS timing board and response box. Each trial began when the subject pressed the START button on the response box, causing the entire text to be displayed with each alpha-numeric character replaced by a dash(–), with normal spacing and punctuation. The participants controlled the word-by-word presentation of the stimuli by pressing the SCROLL button on the response box. Each button press caused the next word of the text to appear and the previous word to be replaced by dashes. Comprehension questions followed the sentences on approximately one-quarter of the trials. Participants responded by pressing YES or NO on the response box and then received feedback as to whether the answer was correct. Participants were instructed to read at a normal pace and carefully enough to answer the questions correctly. The experiment began with seven practice trials and lasted approximately 35 min.

Results and Discussion

Data from two participants who answered fewer than 80% of the questions correctly were replaced with data from additional participants. Mean reading times beginning with *that* and continuing through the auxiliary verb (e.g. *that cheap hotels is/are*) are presented in Table 1. Figure 1 plots the difference in reading times when the singular noun condition is subtracted from the plural noun condition. A positive difference indicates that readers were predominantly influenced by the *that*(Det)-N hypothesis, and a negative difference indicates that readers were predominantly influenced by the *that-as-complementiser* possibility. The word after the noun clearly shows the interaction predicted by the contingent frequency hypothesis: For sentences with an initial *that*, reading times were longer following a plural noun compared to a singular noun, whereas the opposite pattern held when *that* followed a verb.

We performed separate analyses of variance on participant and item reading time means. The factors were list (four lists) or item group (four groups), sentence type (verb followed by *that* or initial-*that*), noun type (singular or plural) and word position (*that*, adjective, noun and auxiliary verb).

The analysis revealed a three-way interaction among sentence type, noun type and word position [$F_1(4,128) = 6.02$, $P < 0.01$; $F_2(4,64) = 9.08$, $P < 0.01$]. The triple interaction is largely due to reading times to the word after the disambiguating noun (the auxiliary verb). At this position, there was an interaction between the number of the noun (noun type) and sentence type [$F_1(1,32) = 18.06$, $P < 0.01$; $F_2(1,16) = 5.86$, $P < 0.05$]. When *that* followed a verb, reading times at the auxiliary verb were longer in the singular noun condition compared to the plural noun condition, though this effect was only reliable by participants [$F_1(1,32) = 6.24$, $P < 0.05$; $F_2(1,16) = 3.04$, $P = 0.10$], indicating that readers initially assumed *that* was a complementiser. The opposite pattern occurred for sentences beginning with *that.* Reading times following the singular noun were shorter than those

TABLE 1
Mean Reading Times (msec) at Specified Word Positions in Experiment 1

	Sentence Region			
Condition	*that*	*cheap*	*hotel(s)*	*is/are*
Post-verb-*that*				
Sg. noun	414	453	484	492
Pl. noun	409	470	479	433
Initial-*that*				
Sg. noun	399	456	454	438
Pl. noun	381	461	479	563

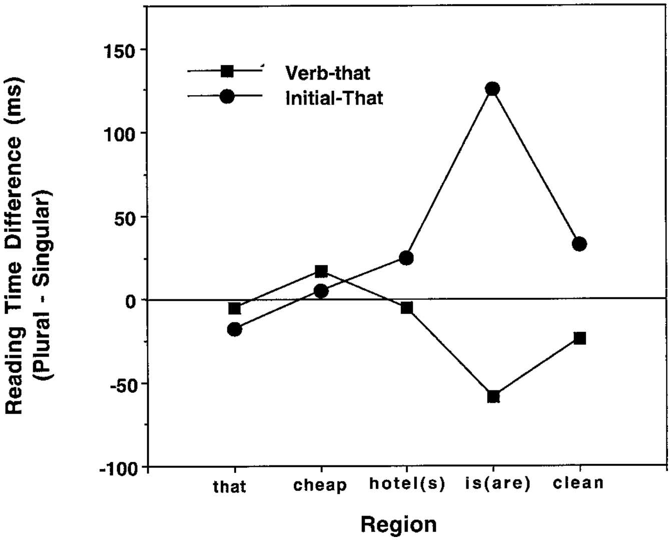


FIG. 1. Reading time differences between the plural and singular conditions in Experiment 1 for the region following the word *that* when it began a sentence (Initial-That) and when it followed the main verb (Verb-that).

after the plural noun [$F_1(1,32) = 22.28$, $P < 0.01$; $F_2(1,16) = 19.26$, $P < 0.01$], indicating that readers initially interpreted *that* as a demonstrative determiner.

These results document the importance of conditional frequency information in language processing. Whereas the absolute probabilities of syntactically ambiguous words are clearly not reliable predictors of processing difficulty across all syntactic environments, grammar-derived conditional probabilities may well be (Jurafsky, 1996). Two-stage models are capable of predicting this general situation because they hold that different sequences of syntactic categories give rise to different first-pass parsing strategies. These results indicate, however, that the choice of the default parse is predictable from statistical properties of the data people learn their language from, rather than from structural principles such as minimal attachment or late closure (Mitchell et al., 1995). Moreover, while a structural simplicity principle such as minimal attachment correctly predicts that a sentence-initial *that*–adjective sequence will be parsed as the beginning of an NP, rather than as a fronted *that* complement, it incorrectly makes the same prediction when a *that*–adjective sequence follows a verb. In the attractor model, statistical differences automatically give rise to different parsing preferences because, in every environment, the most common pattern is associated with the most powerful local attractor.

The statistically based attractor-model is also capable of predicting effects that look very much like structural complexity effects in cases where initial hypotheses need to be revised. Recall that the congruity effect (the difference in reading times when *that* was resolved in its preferred and unpreferred categories) in the initial-*that* condition was about twice the size of the congruity effect in the verb-*that* condition (140 *vs* 65 msec), even though the frequency asymmetry between the more frequent and less frequent reading was larger in the verb-*that* condition (93% complementiser *vs* 6% determiner) compared to the initial-*that* condition (35% determiner *vs* 11% complementiser).

In a symbolic parser, this asymmetry can be explained by taking into account the magnitude of the revision required in each condition. When an initial *that* which was originally taken to be a determiner has to be reanalysed as a complementiser, the noun phrase itself must be reanalysed as the subject of a subordinate clause. In contrast, reanalysis of *that* after a verb from a complementiser to a demonstrative does not affect the overall clause structure. In the attractor model, the asymmetry arises because a big structural revision requires the model to make a large jump across the representation space. This has the effect of it landing in an intermediate location that is relatively far from the nearest attractor, thereby increasing processing time. Later we show the details of how this result obtains in our implemented model.[3]

Experiment 2: Category-contingent Versus Verb-specific Contingent Frequencies

This experiment examined *that* when it followed a verb, comparing NP-bias verbs such as *visit* that do not permit a sentence complement with SC-bias verbs such as *insisted* that are typically used with a sentence complement and do not permit an NP complement. We predicted that the strong category-bias for *that* following a verb to be a complementiser would create a strong attractor that would make processing *that* as a determiner difficult, even when it followed an NP complement verb.

[3] It is possible that there is another reason for the difference in effect-size observed here: the sentential subject construction may be restricted to a relatively erudite speaking/writing style and thus may not occur as frequently in language samples from which typical speakers learn as it does in the Brown corpus, which we have used to estimate relative frequencies. If this is the case, our model will still make the appropriate effect-size prediction if trained on a more natural corpus. But the size difference would then be due simply to a difference in the relative strengths of the complementiser attractor and the determiner attractor, rather than to the revision-difficulty effect just described. It is our suspicion that whatever the natural frequency relationships are, the revision-difficulty effect contributes to the high latency, therefore we have felt it worthwhile to explain how our model captures such effects.

Method

Paricipants. Twenty-eight students from the University of Rochester participated for course credit. All were native speakers of English.

Materials. Four sentences were constructed for each of 20 adjective + noun sequences. Examples using *cheap hotel(s)* are presented in (6):

6a. The lawyer insisted that cheap hotel was clean and comfortable.
6b. The lawyer insisted that cheap hotels were clean and comfortable.
6c. The lawyer visited that cheap hotel to stay for the night.
6d. The lawyer visited those cheap hotels to stay for the night.

Each trial consisted of two sentences, with the second sentence being a natural continuation of the first. On critical trials, the target sentence was always the first sentence in the trial. For the sentence complement verbs, the verb was followed by *that*, an adjective, and then either a singular noun, which disambiguates *that* as a demonstrative in a *that*-less S-complement, or a plural noun, which disambiguates *that* as a complementiser. For the NP-bias verbs, the verb was either followed by *that*, an adjective and a singular noun, disambiguating *that* as a demonstrative, or by *those* followed by an adjective and a plural noun. *Those* was included as a referent-presupposing control. It could be argued that processing difficulty at *that* after an NP-bias verb reflects the fact that it is odd to use a definite determiner without first having introduced a referent. We used the plural *those* rather than the singular *this* as the control because in colloquial speech *this* is often used to introduce a referent despite its definiteness (e.g. *This guy walked in and ordered whiskey*).

Four lists were constructed by assigning each of the four sentences created from a particular adjective + noun sequence to a different list. The target sentences were pseudo-randomly combined with 42 distractor trials, for a total of 62 trials in each list. At least one filler trial appeared before each experimental trial, and conditions were balanced across the two halves of the stimulus lists.

Procedure. The procedure was the same as in Experiment 1.

Results and Discussion

Data from two participants who answered fewer than 80% of the questions correctly were replaced with data from additional participants. Mean reading times beginning with *that* and continuing through the two words that followed the noun (e.g. *that cheap hotels is/are clean*) are presented in Fig. 2.

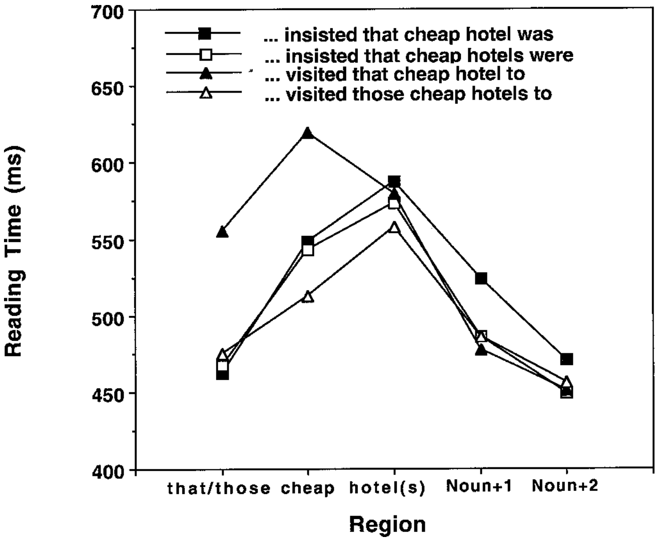


FIG. 2. Mean reading times following NP-bias and SC-bias verbs in Experiment 2.

The results can be understood most clearly by considering the NP-bias verbs and the SC-bias verbs separately. For the NP-bias verbs, reading times were considerably longer to *that* than to *those* at the determiner (*that* or *those*) [$F_1(1,24) = 5.31$, $P < 0.05$; $F_2(1,16) = 5.80$, $P < 0.05$] and at the following adjective [$F_1(1,24) = 23.59$, $P < 0.01$; $F_2(1,16) = 8.12$, $P < 0.05$]. Thus readers had difficulty processing *that* as a demonstrative, demonstrating effects of both the category-level bias and the subcategorisation properties of the specific verb (i.e. *that* after a verb is usually a complementiser and these verbs cannot take a complementiser). The results for the sentence complement verbs replicated those found in Experiment 1. Readers initially interpreted *that* as a complementiser rather than as a demonstrative determiner. Reading times after a singular noun (e.g. at *was*) were longer compared to reading times after a plural noun (e.g. at *were*) [$F_1(1,24) = 4.03$, $P = 0.056$; $F_2(1,16) = 5.16$, $P < 0.03$].

The prediction that reading times to *that* would be longer when it followed an NP-bias verb compared to when it followed an SC-bias verb was clearly confirmed. Reading times at *that* and at the following adjective were clearly longer after an NP-bias verb than after an SC-bias verb [$F_1(1,24) = 8.0$, $P = 0.01$; $F_2(1,16) = 6.15$, $P < 0.03$ at *that*; and $F_1(1,24) = 8.77$, $P < 0.01$; $F_2(1,16) = 2.98$, $P = 0.1$ at the adjective].

The results demonstrated a robust category contingent effect that was not eliminated by verb-specific subcategorisation information. *That* is typically a complementiser when it follows a verb and, consequently, there is a strong bias to interpret it as a complementiser in this environment, even when it follows a verb that cannot take a sentential complement. One could ask, of course, whether these long reading times are due simply to the low rate of occurrence of the word *that* after NP-bias verbs—after all, the point of Experiment 1 was to show that reading times are strongly influenced by context-dependent relative frequencies. However, this hypothesis incorrectly predicts that reading times at *those* following an NP-bias verb would be as long or longer than reading times at *that* in the same environment (the frequency of *those* in this environment is about 30% of the frequency of *that*). Instead, they are significantly shorter. Thus the processing difficulty at *that* cannot be attributed to simple differences in context-dependent expectations. These results pose a problem for any model that simply uses conditional probabilities to assign reading times.[4]

Experiment 3: Contingent Frequencies and Complement Type

The goal of this experiment was to determine whether contingent frequencies also influence phrasal ambiguity resolution. In addition to replicating the NP-bias verb followed by *that* and the SC-bias verb followed by *that* conditions from Experiment 2, we included conditions in which a noun phrase with the determiner *the* (*the*-adjective-noun) followed either an NP-bias or an SC-bias verb.

It is well-established that readers have a strong bias to parse a noun phrase after an NP-V sequence in a main clause as an NP object complement, as in (7a), rather than as the subject of a sentence complement, as in (7b). This bias is reflected in long reading times when the reader encounters the disambiguating verb phrase in a *that*-less sentence complement [e.g. *would cause* in (7b)], which disambiguates the preceding NP (*the angry man*) as the subject of the sentence complement rather than the object of the verb *warned*.

7a. John warned the man about the cheap hotel.
7b. John warned the angry man would cause trouble.
7c. John insisted the angry man would cause trouble.

Now, consider sentence (7c), in which a *that*-less sentence complement follows an SC-bias verb. A number of recent studies have found that readers have little or no processing difficulty at the verb phrase in the complement (Garnsey et al., 1995; Schmauder & Egan, 1995; Trueswell et al., 1993; but cf.

[4] N-gram models (e.g. Brown et al., 1992) and grammar-based conditional probability models (e.g. Jurafsky, 1996) are roughly of this sort.

Ferreira & Henderson, 1990). However, some of these studies have reported increased reading times at the noun phrase (e.g. *the angry man*) compared to noun phrase complements with a *that*, or noun phrases after NP-biased verbs. In a two-stage model, these effects at the NP would be interpreted as rapid revision effects in a system that initially ignores lexically specific structure. However, in an attractor-based system, this difficulty reflects the joint effect of two attractors: the Verb-NP-Det attractor and the Verb-S-Comp-*that* attractor. Thus it is another case of representational intermediacy, due to competing attractors.

Method

Participants. Twenty-eight students from the Univesity of Rochester participated for course credit. All were native speakers of English.

Materials and Procedure. Four sentences were constructed for each of 20 adjective + noun sequences. Examples are presented in (8):

8a. The lawyer insisted that cheap hotel was clean and comfortable.
8b. The lawyer insisted the cheap hotel was clean and comfortable.
8c. The lawyer visited that cheap hotel to stay for the night.
8d. The lawyer visited the cheap hotel to stay for the night.

The fillers, counterbalancing procedure and the experimental procedure were the same as those used in Experiment 2.

Results and Discusssion

Data from four participants who answered fewer than 80% of the questions correctly were replaced with data from additional participants. Mean reading times beginning with *that* or *the* and continuing through the two words that followed the noun (e.g. *that cheap hotel is clean*) are presented in Fig. 3.

There was a three-way interaction between the type of verb (SC-bias or NP-bias), type of determiner (*that* or *the*) and the position of the word [$F_1(3,72) = 4.50$, $P < 0.01$; $F_2(5,80) = 4.32$, $P < 0.01$]. This result can be understood most clearly by considering several crucial comparisons. Reading times following *that* were again longer following NP-bias verbs than SC-bias verbs, replicating the pattern found in Experiment 2. In this experiment, however, the effect did not show up until the adjective [$F_1(1,24) = 9.84$, $P < 0.01$; $F_2(1,16) = 18.25$, $P < 0.01$].

Reading times were longer at the (singular) noun after an SC-bias verb with a *that* compared to an SC-bias verb with a *the* [$F_1(1,24) = 6.89$, $P < 0.02$; $F_2(1,16) = 4.10$, $P = 0.06$], and at the auxiliary verb [$F_1(1,24) = 6.55$, $P < 0.02$; $F_2(1,16) = 16.22$, $P < 0.01$]. This result indicates that readers had a

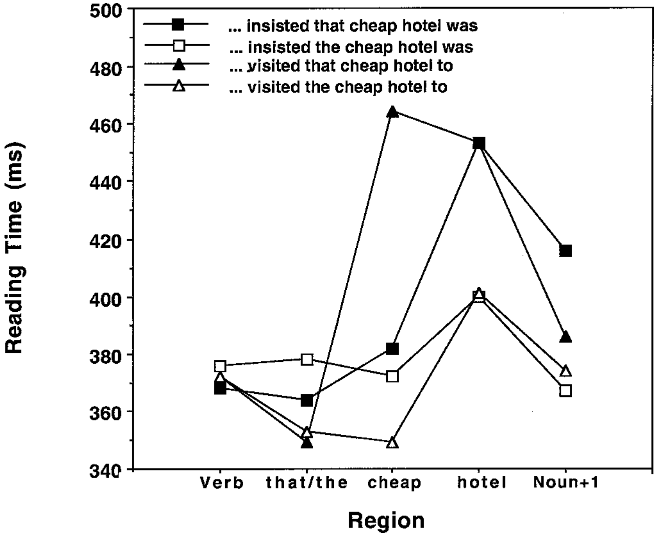


FIG. 3. Mean reading times following NP-bias and SC-bias verbs in Experiment 3.

bias to interpret *that* as a demonstrative until they encountered the singular noun, replicating the result found in Experiment 1. The full timecourse of the *that* effect following NP-bias verbs can be seen by comparing the NP-bias *that* and the NP-bias *the* conditions. Reading times in these conditions diverge at the adjective and converge after the noun.

Reading times were 25 msec longer at the determiner following SC-bias verbs than NP-bias verbs [$F_1(1,24) = 5.53$, $P < 0.05$; $F_2(1,16) = 2.23$, $P = 0.15$], and 20 msec longer at the adjective [$F_1(1,16) = 3.06$, $P < 0.10$; $F_2(1,16) = 2.23$, $P = 0.19$]. Although these results are only marginally reliable, a similar pattern was reported by Trueswell et al. (1993) and Garnsey et al. (1995). Furthermore, Experiment 4 replicates this effect with a larger sample of verbs, as will be seen later. The reason for the high variability across items will become clearer once we begin to explore the details of the attractors for this environment constructed by our simulations.

In sum, the results of this experiment further illustrate the interactions among attractors that we observed in Experiment 2. Reading times were again longer in the *that* condition following NP-bias verbs than SC-bias verbs, replicating the effect observed in Experiment 2. In addition, reading times were longer at *the* and the following adjective after SC-bias verbs compared to NP-bias verbs. While this effect was quite small, it was predicted to arise from competition between the NP-Det-object attractor and the Sbar-comp attractor at *the*. NP-bias verbs followed by *the* are easy to process because all the evidence favours placing the processor near the NP-object attractor. But in cases where an SC-bias verb is followed by *the*, the evidence from the verb favours the sentence complement attractor and the evidence from the word *the* favours the NP-object attractor. Again, the processor ends up in an intermediate position, far from both attractors, and thus gravitation time is high. In the next section, we diagram the attractors and show the relevant intermediate locations in our implemented model.

Summary of Contingent Frequency Effects

Four contingent (conditional) frequency results emerged from Experiments 1–3:

1. *That* at the beginning of a sentence was initially taken to be a demonstrative determiner, whereas *that* after a verb was taken to be a complementiser. These biases reflect the contingent frequencies in the language, as determined by corpus analyses, and were interpreted as being due to the relative strength of the *that*-complement and *that*-Det attractors in the two environments.
2. The difficulty of revising the interpretation of a sentence-initial *that* from a determiner to a complementiser was greater than revising the interpretation of a post-verbal *that* from a complementiser to a determiner. We hypothesised that this asymmetry arises because bigger structural revisions generally require the model to make large jumps across the representation space, resulting in it landing in an intermediate location that is relatively far from the nearest attractor.
3. Readers experienced processing difficulty when *that* followed a verb that did not permit a complementiser. Thus the category-based bias for *that* as a complementiser when it followed a verb still exerted large effects even when it ran counter to unambiguous verb-specific information. This was hypothesised to be a result of the strong effects of a *that*-complement attractor after a verb.
4. There was a bias to parse an NP beginning with the determiner *the* following a verb as an NP complement. This bias caused processing difficulty for NPs that followed even those verbs that nearly always occur with sentence complements. These effects were attributed to the strength of the NP-Det-object attractor.

It is important to note that the contingent frequency effects that are hypothesised to arise from the influences of attractors cut across different linguistic levels or grains. For example, results 1 and 3 show that in resolving a lexical category ambiguity, the processing system is sensitive to category-contingent frequencies. Result 4 shows that phrasal ambiguities are also sensitive to contingent frequencies.

IMPLEMENTING A DYNAMICAL SYSTEM

We earlier outlined how concepts from dynamical systems theory can be used to model the general correlation between ambiguity and increases in reading times. The central observation was that a dynamical system is dominated by the properties of a single attractor when it is sufficiently near to it, but it can show the simultaneous influences of several attractors when it is in a more intermediate location. This motivated letting attractors correspond to syntactic categorisations and letting different cases of lexical items in context vary according to whether they put a dynamical processing system close to one attractor or intermediate between several. We now describe an explicit model implementing these ideas and show how the model generates the pattern of the reading time results presented in Experiments 1–3.

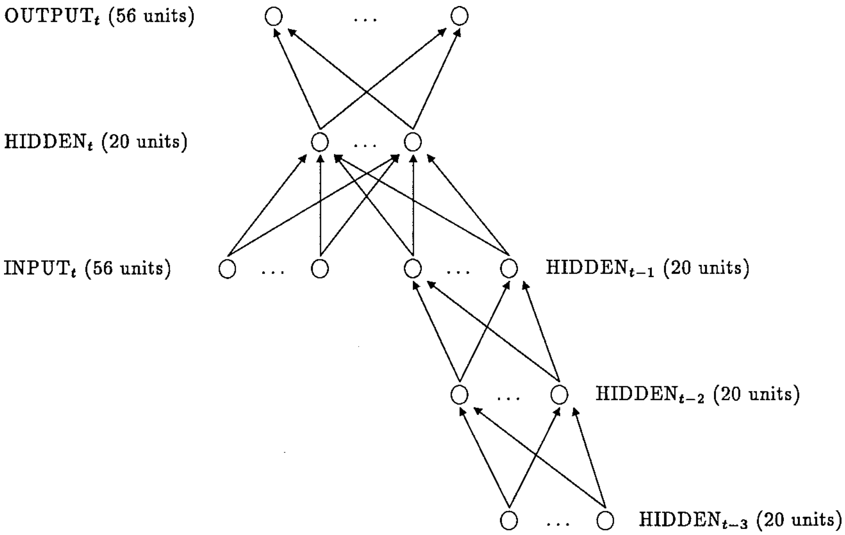
Implementation

Implementing the model requires: (1) creating a similarity-based representational space with appropriate attractors that places the input somewhere in the representational space, and (2) implementing an algorithm that maps the attractor dynamics onto reading time.

The representational foundation of our dynamical model is its similarity-based representation space. We define a *visitation set* to be a set of points in the representation space visited by the processor while it processes a large corpus approximating natural usage.

Network Architecture

To create empirically testable representation spaces and visitation sets, we used the augmented simple recurrent network shown in Fig. 4. This network has exactly the same relaxation dynamics as the architecture of Elman's (1990, 1991) syntactic structure experiments, in which there was only one copy of the hidden layer. But there is a slight difference with Elman's device in the learning: at each time step, error signals from all pairs of successive hidden layers are added to the error signal under backpropagation. This is the technique known as "backpropagation through time" described in Rumelhart, Hinton and Williams (1986) and Pearlmutter (1995), with the proviso that only three prior time steps of hidden unit activation were taken into account. Summing error signals over three pairs


OUTPUT$_t$ (56 units)
HIDDEN$_t$ (20 units)
INPUT$_t$ (56 units)
HIDDEN$_{t-1}$ (20 units)
HIDDEN$_{t-2}$ (20 units)
HIDDEN$_{t-3}$ (20 units)

Activation functions:

Hidden layer t:

$$f(\text{net}_i) = \frac{1}{1 + e^{\text{net}_i}}$$

Hidden layers $t-1$, $t-2$, $t-3$:

> The activation pattern on hidden layer $l-1$ is the same as the activation pattern that appeared on hidden layer l when the previous input was presented (for $l \in \{t, t-1, t-2\}$).

Output units:

$$f(\text{net}_i) = \frac{e^{\text{net}_i}}{\Sigma_k \, e^{\text{net}_k}}$$

where k ranges over output units and

$$\text{net}_i = b_i + \sum_j w_{ij} a_j$$

where b_i is the bias on unit i, a_j is the activation of unit j, and w_{ij} is the weight from unit j to unit i.

Cost function:

$$C = \sum_k t_k ln \; a_k$$

where k indexes output units, a_k is the activation of unit k, and t_k is the target for unit k. The weights were adjusted every time a new word was presented to the network.

FIG. 4. The recurrent connectionist network that forms the core of the attractor-based model.

of hidden unit layers instead of just one (as in Elman's learning procedure) makes it a little easier for the network to learn long-distance dependencies, but does not otherwise seem to change the types of representations discovered by the network. It would, in fact, be necessary to sum error signals over all previous time steps to be certain of the convergence results normally associated with the backpropagation algorithm. The method employed here, using only a few recent time steps, sometimes referred to as "truncation of the gradient", is not guaranteed to converge, but it always achieved stability in the current simulations. For purposes of clarity, we will only use the terms "SRN" and "simple recurrent network" to refer to networks with architectures like those of Elman (1990, 1991), in which backpropagation was carried out through only one prior time step.

As in Elman's model, the inputs were always taken to be localist representations of words: for each word, one unique bit was on, all the other inputs were off. The words were taken in sequence from a corpus. The network's task was to predict on the output layer what word was coming next at each point. Since the optimal target outputs under this set-up form a probability distribution, the output units employed the softmax activation function. The hidden units employed the fixed sigmoid activation function. The weights of the network were adjusted by backpropagation of the error signal through the network for the feedforward connections and through time for the recurrent connections (see Rumelhart et al., 1986). In the three simulations described below, we used the approximation of the error gradient that resulted from truncating the recurrent backpropagation after a fixed number of time steps (see Pearlmuter, 1995, for a discussion).

To generate training data for the models described in this section, we used probabilistic finite-state grammars which approximated those properties of English that were relevant in each experiment. We trained the network model on the output of the grammar until it seemed qualitatively to have learned all the distinctions encoded by the grammar and its average error had come near an asymptote.

The Dynamical Processor

We used the network to define a dynamical system for sentence processing. First, we created a visitation set by using the training grammar to generate a sample corpus of n words, feeding this corpus to the network with learning turned off, and recording the set of n hidden unit locations visited during the presentation of the sample. We interpreted this set of n points as a set of fixed bodies, all of equal mass, at locations $\vec{x}_1 \ldots \vec{x}_n$, and defined the following dynamical behaviour for a movable test body at any point, $\vec{x}$, in the space:

$$\frac{\mathrm{d}\vec{x}}{\mathrm{d}t} = m \sum_{i=1}^{n} \frac{(\vec{x}_i - \vec{x})}{r_i^p}$$

where r is the length of $(\vec{x}_i - \vec{x})$ and m is the mass of each fixed body.

The position of the movable test body at any time corresponds to the state of the processor. Each time the system gets a new word, the test body is placed in a new starting position in the representation space and it must gravitate to an attractor before the next word can be processed. When $p = 2$, this system is like a Newtonian gravitational system with a test body of mass 1 and velocity equal to 0 at infinite distance from the origin. The value of m determines how quickly the system travels along its trajectories, but varying it over the positive real numbers does not qualitatively affect the behaviour of the system. The value of n does affect the behaviour of the system in the sense that if one samples the space too sparsely, then the cluster structure does not reflect the structure of the grammar. It seemed to be necessary to set n to a higher value in the simulation of Experiment 1 than in the simulation of Experiments 2 and 3. We do not know why this is at present. The value of p determines the number of attractors in the system: if $p = 0$, then the system has only one attractor located at the centre of mass of all the fixed bodies; if p is greater than 0, there are multiple attractors, one at each fixed body, and there is a complex network of saddle points located at the centres of masses of dense clusters.[5] The implemented system (see below) is sensitive to the value of p in an important way and so we had to tune this parameter carefully.

To implement this dynamical system in a practical fashion, we made several simplifications. The attractors defined by equation (1) are singular points—the velocity of the test body goes to infinity at them and gets very large near them. To smooth out these singularities, we replaced r with r_{min} whenever r was less than r_{min}.[6] This put a cap on the influence that any one fixed body could have on the test body. It also had the effect of transforming the saddle points at cluster means into attractors. Thus attractors corresponded to behaviours that were classified as similar by the SRN. Note, however, that "cluster" and "classified as similar" are not precise terms. Indeed, the configuration of attractors in this implemented system varied as the value of p was changed. Crudely speaking, a smaller value of p implies fewer attractors and a larger value of p implies more attractors. We should note the relationship between p and the number of attractors is non-linear and there are ranges in which changing the value of p has no effect on the

[5] Saddle points are fixed points which attract trajectories from some regions of the space and repel them into others.

[6] This is roughly like assuming that the test body can never get any closer to a fixed body than r_{min}, although it does not treat the complex interactions at close range of bodies composed of real matter.

topology of the system. In other words, this variation is typical of bifurcation profiles in complex dynamical systems (see Strogatz, 1994, for an introduction).

The exponent p can thus be viewed as a "grain size" parameter, related to the grain size discussed by Mitchell et al. (1995). We adjusted this parameter in each simulation to make the set of attractors line up nearly perfectly with the set of states distinguished by the grammar. This important flexibility reveals a certain stipulativeness of the model, but it is significant that the correspondence was near perfect in the two simulations in which the states distinguished by the grammar were easily identified [the simulations of Experiments 1 and (2 and 3)], since there are many computation systems which the network + gravitation model cannot emulate no matter what value of p is used. With p set to achieve this near perfect line-up, we nevertheless saw influences of constraints at different traditional grains (lexical and phrasal), so the model achieved the desirable result of letting these levels interact without losing predictive power.

We used the simplest integration technique (i.e. Euler's method) to find approximations of the system's trajectories: take the change in the system's state at x to be $\Delta t(\mathrm{d}\vec{x}/\mathrm{d}t)$ for Δt, a small positive constant.[7] For simplicity, we fold the two proportionality constants, m and Δt, into one, $\mu = m\Delta t$ in the remaining discussion. The revised system is shown in equation (2):

$$\Delta\vec{x} = \mu \sum_{i=1}^{n} \frac{(\vec{x}_i - \vec{x})}{r_i^p}$$

Since the attractors of this system can be either complex cycles or stable fixed points, we detected them by looking for places where either the trajectory changed direction by more than 90° in one time step or where the velocity slowed to below a small value.[8] To model reading times for a sentence S consisting of words $w_\mathrm{i}, w_2, \ldots, w_k$, we presented S one word at a time to the RCN and recorded the hidden unit locations $\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_k$ associated with each word presentation. The predicted reading time at word w_k was then taken to be the number of time steps it took the system described in equation (2) to gravitate from $\vec{h}_k$ to an attractor.

Simulation of Experiment 1

We now describe a simulation of the central results of Experiment 1.

[7] We chose not to use a more exact integration technique like 4th order Runge-Kutta because such a method was, computationally, extremely costly and seemed to make little difference in the outcomes that concern us here.

[8] It turned out to work well to use r_{min} as this small value.

TABLE 2
Corpus-generating Grammar for Experiment 1 Simulation

0.90	S : Det[sg] N[sg] VP[sg] p
0.10	S : that N[pl] V[0][pl] V[AdjP][sg] AdjP p
0.60	VP[sg] : V[0][sg]
0.30	VP[sg] : V[S′][sg] that N[pl] V[0][pl]
0.10	VP[sg] : V[S′][sg] Det[sg] N[sg] V[0][sg]
[Zipf]	V[0][sg] : 0.44 sings, 0.22 stops, 0.14 talks, 0.11 whistles, 0.09 leaps
[Zipf]	V[0][pl] : 0.44 sing, 0.22 stop, 0.14 talk, 0.11 whistle, 0.09 leap
[Zipf]	V[S′][sg] : 0.34 thinks, 0.17 agrees, 0.11 insists, 0.09 wishes, 0.07 hopes, 0.06 remarks, 0.05 pleads, 0.04 speculates, 0.04 doubts, 0.03 hints
[Zipf]	V[AdjP][sg] : 0.44 is, 0.22 seems, 0.14 looks, 0.11 sounds, 0.09 becomes
[Zipf]	AdjP : 0.44 funny, 0.22 strange, 0.14 good, 0.11 tolerable, 0.09 surprising
[Zipf]	Det[sg] : 0.44 the, 0.22 that, 0.14 a, 0.11 this, 0.09 one
[Zipf]	N[sg] : 0.34 woman, 0.17 man, 0.11 dog, 0.09 cat, 0.07 marmot, 0.06 girl, 0.05 boy, 0.04 artist, 0.04 journalist, 0.03 sailor
[Zipf]	N[pl] : 0.34 women, 0.17 men, 0.11 dogs, 0.09 cats, 0.07 marmots, 0.06 girls, 0.05 boys, 0.04 artists, 0.04 journalists, 0.03 sailors

Training the Network

The grammar used in the simulation of Experiment 1 is shown in Table 2. The grammar generated the word *that* in the four different kinds of environments that were the focus of Experiment 1: determiner versus complementiser in sentence-initial position and determiner versus complementiser in post-verbal position. The relative frequencies within these environments did not correspond directly to estimates of frequencies in natural usage, but they were biased in the same direction relative to 50% (see Table 3). We chose simulation values that were less extreme than the corpus values to facilitate training the network. Later, we explore a network trained from an actual natural language corpus with more representative frequencies.

The relative frequencies of different lexical items within classes were set according to "Zipf's Law", which holds that a rank versus frequency plot of

TABLE 3
Comparison of Relative Frequencies in the Brown Corpus and Simulation 1 Grammar
(Brown Corpus Statistics from the Penn Treebank)

Type of that	*Brown Corpus*	*Simulation Grammar*
Sentence-initial compl. *vs* det.	25%	34%
Post-verbal compl. *vs* det.	96%	93%

the vocabulary elements drawn from any large natural language corpus forms the cusp of a hyperbola (Zipf, 1949).[9] The law has been confirmed by Zipf and his successors as a fair approximation for numerous corpora in a wide range of languages. We have also observed that it gives a reasonable approximation within several of the main lexical categories in the Brown corpus (e.g. noun, verb, adjective, adverb, determiner) and therefore used it to assign frequencies to lexical items in the grammar.[10]

In the grammar used for simulating Experiment 1, there was a perfect correlation between whether *that* was used as a complementiser or determiner and whether it was followed by a plural noun or a singular noun. Thus the grammar generated sentences like *That dogs bark is undeniable* and *Josiah knows that dogs bark*, and sentences like *That dog barks* and *Josiah knows that dog barks*. However, it never generated sentences of the form *That a dog barks is undeniable* or *Josiah knows that a dog barks*. We decided not to include the latter types of sentences because recurrent networks have difficulty learning long-distance dependencies, without special training procedures (e.g. Elman, 1993). Creating a categorical relationship between the syntactic category of *that* and the number of the subsequent noun made it easier for the network to learn the long-distance dependency between the way *that* is used locally and the type of environment its clause occurs in. These unrealistic assumptions about the corpus may introduce some biases in the simulation which distort its correlation with the reading time data (i.e. the network gets unequivocal evidence that sentence initial *that* is a complementiser when it encounters a plural noun, whereas people do not). However, this distortion should not interfere with the expected effect of frequency contrasts on reading times.

Reading Time Predictions

To train the network for Experiment 1, we generated sentences at random using the Experiment 1 grammar and fed them to the network one word at a time. Thus there was not a finite training set but an open-ended sequence, and it was not useful to evaluate the total error over all training examples to determine the network's performance. For the simulations of Experiments

[9] In other words, if one orders (ranks) the vocabulary elements by their frequencies and then numbers them sequentially with the integers 1, 2, 3, and so forth, a plot of these integers versus the corresponding frequencies forms a hyperbola. This implies that if one plots the log of rank versus the log of frequency, the result should be a straight line.

[10] The Brown corpus rank-frequency plots of lexical classes seem, in fact, to be slightly biased in favour of lower frequency elements (i.e. the highest frequency elements seem to be missing from the tabulation), so a log-log plot is not linear but bowed slightly upward in the middle. We have found that distortions of this sort and even more radical divergence make little difference in the qualitative outcome of the simulation.

1, 2 and 3, the training grammar was a finite-state grammar with only a few states. Therefore, we were able to determine when the network was making all the right distinctions simply by inspecting its outputs in all the relevant conditions. We also measured error quantitatively as a weighted sum over current and preceding pattern errors [Error(t) = 0.99*Error(t − 1) + 0.01*$|C|$, where t indexes the time step and C is the error on the current pattern (see Fig. 4)] and made sure that this error asymptoted before we stopped training.

We used a learning rate of 0.03. For four out of four initial weight configurations, the Experiment 1 network learned all the distinctions in the training grammar by the time at least 400,000 words had been presented. In fact, it typically learned the distinctions in less than 100,000 presentations, but we ran the training longer to allow the hidden unit clusters to become a bit more distinct. We studied the representations for all four training episodes and found them to be essentially similar. The results reported here are based on one typical episode.

After a small amount of experimentation with parameter values, we set $n = 6000$, $p = 3.5$, $r_{\min} = 0.01$ and $\mu = 0.00005$ for the simulation of Experiment 1. Figures 5 and 6 compare the human reading time data in Experiment 1 to the simulation results. The pattern of reading times generated by the model is similar to that observed in Experiment 1. The main difference is that the bulk of the processing difficulty occurs about a word later in the human data than in the simulation results. This may be because effects are often delayed by a word or two in self-paced reading. It also may be that the unrealistic simplicity of the training grammar contributes to the effect. For example, the fact that the number marking on a noun following the word *that* gives a categorical signal in the grammar as to the relation of the noun to the word *that*, means that the model encounters clear, unexpected information earlier than humans do (after all "That marmots . . ." might turn out to be "That marmots' den . . ." in English but not in the simulation grammar). These effects may weaken the synchrony of human and model results, but the source of the reading time contrasts is still plausibly the same in humans and in the model.

Why does the model make the appropriate predictions? The gravitational dynamical system for this simulation has three main attractive regions, corresponding to the three major types of categories represented in the grammar: verb, determiner and noun. Each of these regions contains several distinct attractors, which correspond to the different syntactic roles the words in each category can play. Since the effect of interest shows up on the nouns in this simulation, it is revealing to take a close-up look at the noun region.

Figure 7 shows (as black dots) the fixed bodies in the noun region in two dimensions. These dimensions are the first and second principal

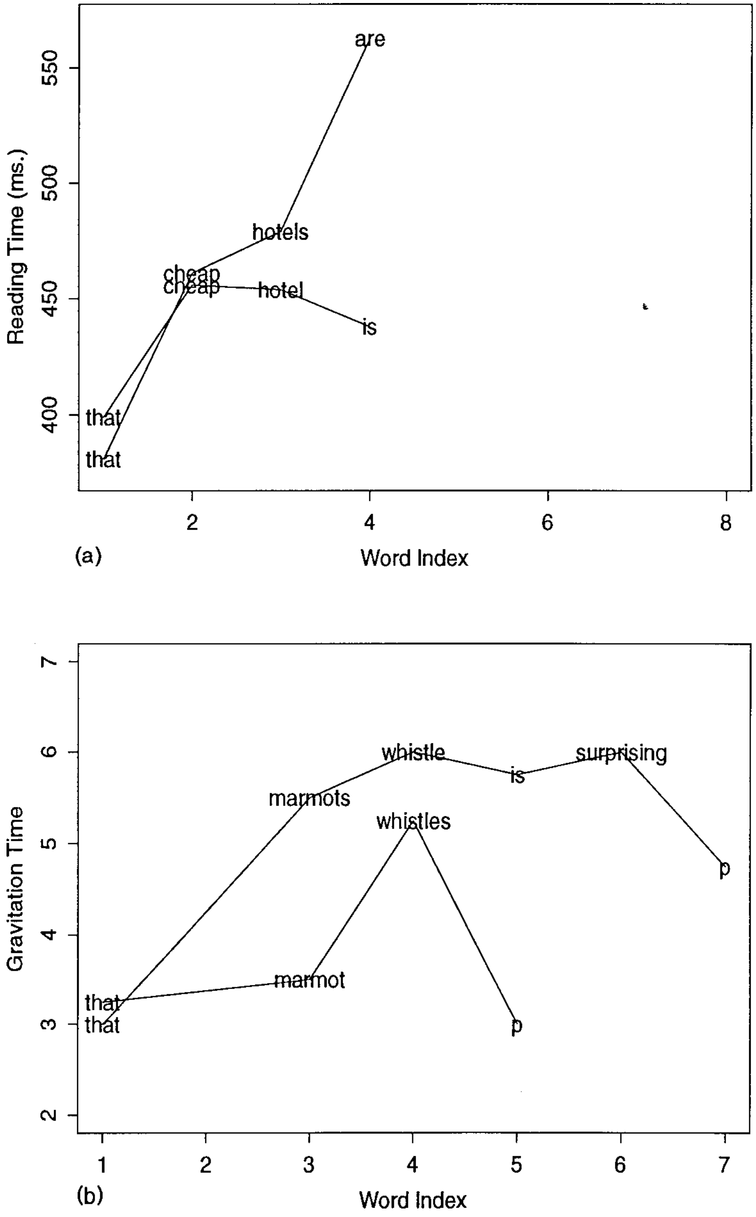


FIG. 5. Comparison of human participant results (a) and simulation results (b) for Experiment 1, initial *that* condition.

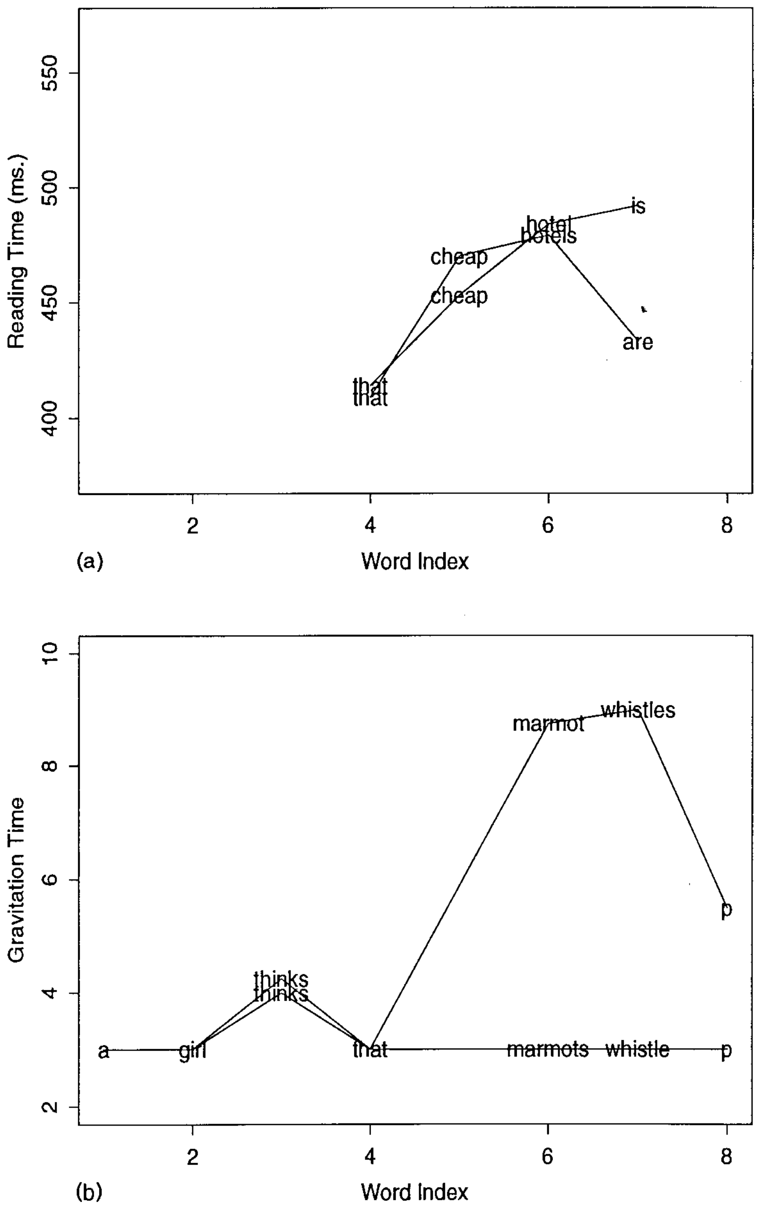


FIG. 6. Comparison of human participant results (a) and simulation results (b) for Experiment 1, post-verbal *that* condition.

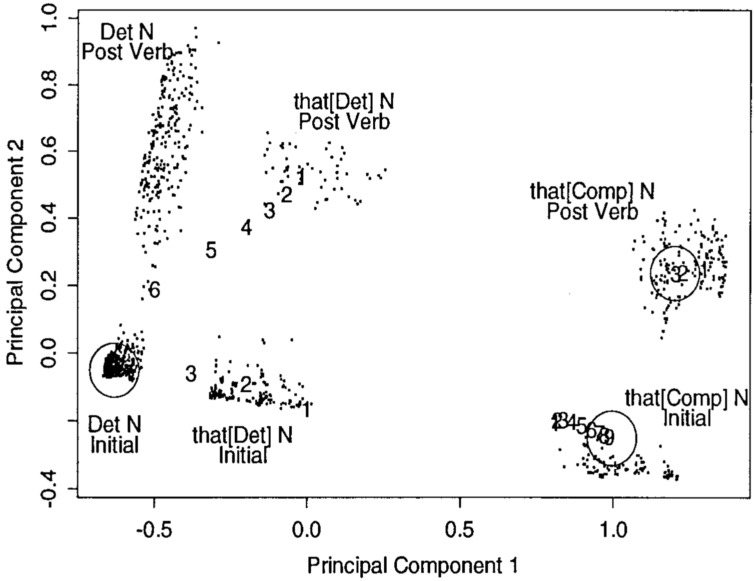


FIG. 7. A portion of the visitation set for simulation 1 with labelled clusters representing nouns in different syntactic environments.

components, respectively, of the points corresponding to nouns.[11] Clusters of points within this region correspond to distinct grammatical situations and are labelled accordingly. For example, the cloud of points labelled "that[Det] N, Initial" corresponds to nouns that occurred after the word *that* sentence-initially, where *that* was functioning as a determiner (thus the noun was singular). The cloud of points labelled "Det N, Initial" corresponds to nouns that occurred after some determiner other than *that* in sentence-initial position. The circles identify the three attractors in the noun region, which may be characterised as subject of an embedded sentence ("that[Comp] N, Post Verb"), subject of a sentential subject ("that[Comp] N, Initial"), and head of the subject of the matrix sentence ("Det N, Initial").[12] If the test body of the dynamical system is placed anywhere in the depicted region, it will gravitate to one of the three circled attractors. The sequences of numbers in the diagram show sample trajectories of the test body: each

[11] The first two components account for 80 and 15% of the variance in the noun subportion of the visitation set, respectively, so Fig. 7 shows nearly all of the structure of this subset.

[12] Note that the generating grammar contains only subject nouns, no object nouns or nouns in other syntactic positions.

trajectory starts at the point labelled "1" and proceeds in order of the sequence, ending finally at the attractor towards which it is heading.[13]

The three attractors in Fig. 7 correspond to the important syntactic distinctions that the grammar makes between nouns: there can be statistical differences between the behaviours of nouns within the same attractor basin, but these tend towards zero as larger and larger samples of the language are examined. In this sense, the dynamical model makes explicit the "emergent" organisation of the neural network representation which gives rise to syntactic constraints on processing. Moreover, although the syntactic distinctions between nouns made here are not the same ones that a linguist would choose for parsing the real language, English, if a linguist were studying the output of the training grammar, then the distinctions formed by the dynamical model would probably be the distinctions of choice.

The cluster arrangement in this region is robust under repeated training episodes with different random starting weights in the RCN. In four successive simulations, the same pattern appeared. In fact, the locations of the clusters are easy to predict based on the principle that locally ambiguous constructions receive intermediate representations: there are three clusters corresponding to the three distinct noun-states determined by the grammar—these are the clusters corresponding to the three attractors just mentioned. The remaining clusters correspond to non-distinct states which share features with other states: the "that[Det] N, Initial" cluster is near the subject determiner attractor but is displaced in the direction of the sentential subject region; the "Det N, Post Verb" cluster is near the subject determiner attractor but is displaced in the direction of other post-verbal structures; the "that[Det] N, Post Verb" cluster is near the subject determiner attractor but is displaced simultaneously in the direction of other post-verbal structures and in the direction of the embedded sentence structure (also introduced by the word *that*).

We can see from the trajectories shown in Fig. 7 how the model makes appropriate predictions. A singular noun following sentence-initial *that* ("that[Det] N, Initial") is processed quickly because it starts from a cluster right next to the powerful matrix subject attractor, which rapidly pulls it in. The matrix subject attractor is powerful because of the high frequency of matrix subjects. A plural noun following sentence-initial *that* ("that[Comp] N, Initial") takes longer to process because it lands near the much weaker sentential subject attractor. The sentential subject attractor is weaker because of the comparatively low frequency of sentential subjects. This contrast gives rise to the gravitation time difference at the noun shown in Fig. 5.

[13] The parts of the trajectory which are nearly at the attractor are not labelled so as to avoid confusing overlaps of labels.

On the other hand, a plural noun following post-verbal *that* ("that[Comp] N, Post Verb") is processed reasonably quickly because its trajectory starts very near the embedded subject attractor. By contrast, a singular noun following post-verbal *that* ("that[Det] N, Post Verb") takes somewhat longer to process, even though it is drawn in by the powerful subject attractor, because it starts quite far away from this attractor.

Two features of the dynamical system contribute to the correct predictions: (1) A strong attractor pulls the test body more quickly than a weak attractor—this predicts the general correlation between the frequency of a construction and its processing ease. (2) Gravitation takes longer if the starting point of a trajectory is close to an attractor rather than far away—this gives rise to the inverse correlation between the ambiguity of a construction and its processing ease. Thus the attractors both play the role of traditional categorical sentence fragment parses, and also interact to make appropriate predictions about the roles of frequency and similarity which constraint-based models have emphasised. Since the attractors develop through the learning process of the RCN, they are appropriately thought of as the "emergent properties" which constraint-based researchers have hypothesised to account for syntactic effects within a lexicalist framework (e.g. Juliano & Tanenhaus, 1994). It is important to note that attractors corresponding to higher level phrasal categories emerge even though the system is lexicalist in nature. This is important given that critics of both lexicalist and constraint-based models have often appealed to category-level biases that precede lexical heads as evidence against these approaches. Our results show that this type of argument is misleading.

The fact that the emergent properties are related to the learning process of the RCN raises the question of why we did not simply choose to work with the dynamical properties of the RCN. The RCN itself is a dynamical system in two senses: the recurrent connections in the hidden layer give rise to dynamical structures (cf. Rodriguez, 1995; Wiles and Elman, 1995) and the learning process is dynamical. We have chosen to adopt a separate dynamical system to interpret the representation adopted by the RCN because it seemed like the best way to capture the key conceptual generalisations we observe in processing. We chose not to work with the dynamical structures associated with the recurrent connections of the RCN because we did not find these useful for our purposes (a more detailed discussion of this point follows later). We also chose not to work directly with the dynamics of learning. We believe, however, that there may be an interpretation of the dynamics of learning which corresponds closely to the interpretation we create using the gravitational model: there are locations in the representation space associated with important categorical distinctions, and the representations assigned to specific items migrate towards these locations during the course of training. However, it is difficult to analyse these properties of the learning dynamics because the locations of the attractors are constantly changing. Further analysis of learning in this light may ultimately be helpful and we see this project as laying a foundation for such work.

Simulations of Experiments 2 and 3

Since the important grammatical environments in Experiments 2 and 3 did not require learning of the long-distance dependencies involved in Experiment 1, we used a more repesentative grammar, presented in Table 4. The network had 36 inputs, 5 hidden units and 36 outputs, and 3 time steps were unfolded in time during learning (Pearlmutter, 1995; Rumelhart et al., 1986).

The grammar for simulations 2 and 3 encodes the relevant overall bias in favour of assigning a complementiser interpretation in V + *that* sequences (Table 5, Line 1). The grammar assumes that the complementiser is present (i.e. not deleted) after about 70% of verbs introducing a sentence complement. This corresponds to the pattern found in the Treebank corpus once one eliminates sentence complements with nominative pronouns, which typically occur without complementisers (Table 5, Line 3). This gives

TABLE 4
Corpus-generating Grammar for Experiments 2 and 3 Simulation

1.00	Sroot : S p
1.00	S : NP VP
0.67	VP : VP[NP]
0.33	VP : VP[S′]
0.67	VP[NP] : V[NP] NP
0.33	VP[NP] : V[NP]
1.00	VP[S′] : V[S′] S′
0.67	S′ : that S
0.33	S′ : S
1.00	NP : Det N
[Zipf]	Det : 0.44 the, 0.22 a, 0.14 which, 0.10 that, 0.10 those
[Zipf]	V[NP] : 0.34 called, 0.17 followed, 0.11 pulled, 0.09 caught, 0.07 pushed, 0.06 loved, 0.05 visited, 0.04 studied, 0.04 tossed, 0.03 grabbed
[Zipf]	V[S′] : 0.34 thought, 0.17 agreed, 0.11 insisted, 0.09 wished, 0.07 hoped, 0.06 remarked, 0.05 pleaded, 0.04 speculated, 0.04 doubted, 0.03 hinted
[Zipf]	N : 0.34 woman, 0.17 man, 0.11 dog, 0.09 cat, 0.07 blouse, 0.06 hat, 0.05 cake, 0.04 ball, 0.04 watch, 0.03 cypress

TABLE 5
Comparison of Relative Frequencies in the Brown Corpus and the Training Grammar for Simulations 2 and 3 (Brown Corpus Statistics from the Penn Treebank)

	Brown Corpus	*Simulation Grammar*
1. Complementiser interpretation of V + *that*	93%	80%
2. Relative frequency of NP complements *vs* sentential complements	79%	67%
3. Relative frequency of *that*-deletion in sentential complements (not counting cases in which the embedded subject is a nominative pronoun)	<50%	33%

rise to a strong complementiser attractor, which plays an important role in Experiment 2. The grammar also makes the realistic assumption that there is a high frequency of noun phrase complements in comparison to sentence complements (Table 5, Line 2). This gives rise to a strong direct object attractor, which plays a central role in Experiment 3. The grammar includes an option of using an NP-verb intransitively. This is not unrealistic, but in this case, the main point of including this feature was to facilitate learning the distinction between subjects and objects.

To train the network for simulations 2 and 3, we generated sentences at random using the Experiment 2/3 grammar and fed them to the network one word at a time. Again, the training grammar is a finite-state grammar with only a few states, so we were able to determine when the network had achieved satisfactory performance by inspection. We also checked to make sure the error measure described on p. 237 had asymptoted.

We used a learning rate of 0.03. For four out of five initial weight configurations, the network learned all the distinctions in the training grammar by the time at least 100,000 words had been presented. This grammar was a little harder to learn than the Experiment 1 grammar because of the dependency involved in making the subject/object distinction, and when the network failed, it was always because the network failed to clearly distinguish subjects and objects. We trained several of the successful networks up to the 400,000th word presentation to achieve tighter clustering of the hidden unit representations. In every such case we examined, a similar representation pattern emerged. The results reported here are based on one typical case.

For these experiments, we set $n = 2000$, $p = 2.4$, $r_{\min} = 0.01$ and $\mu = 0.0002$ in the gravitational model. In this case, the parameter p needed to be fine-tuned to make a crucial distinction, which we will discuss shortly.

Figure 8 provides a comparison between the reading time data and the simulation results for the contrast between *those* and *that* and following an NP-complement verb (Experiment 2). Figure 9 compares the human and

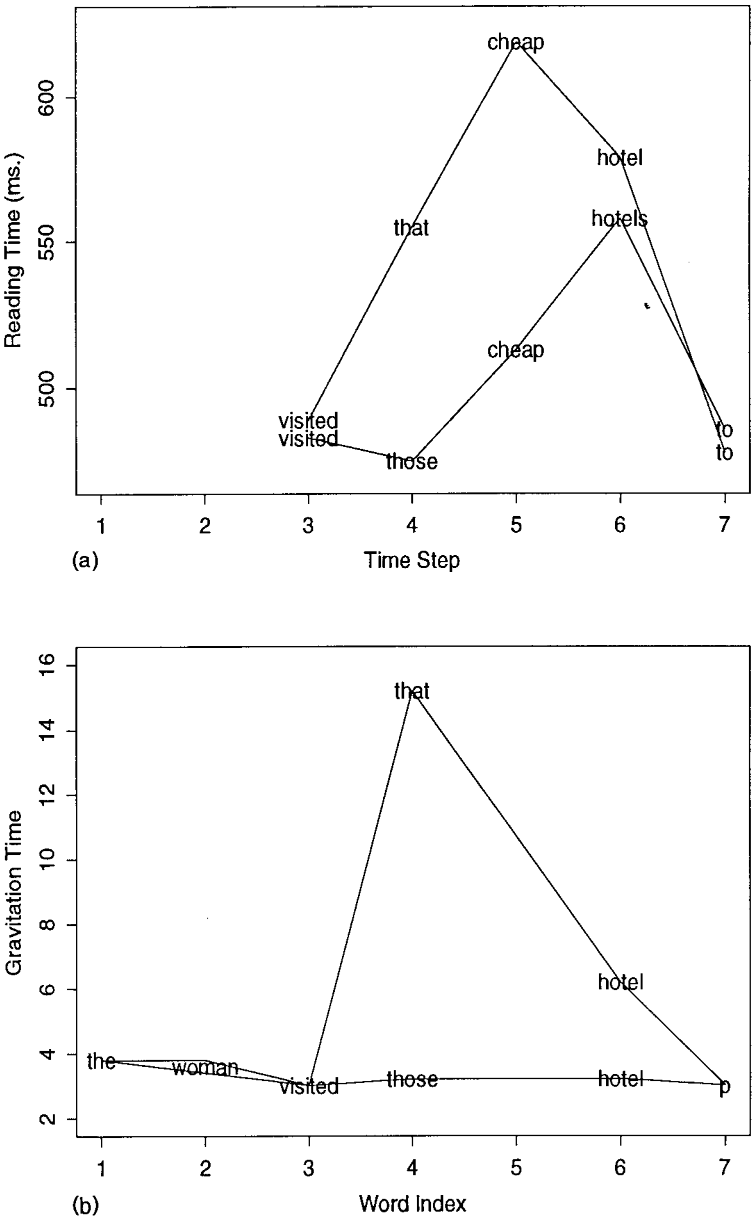


FIG. 8. Reading time at *that* compared with reading time at *those* following an NP-only verb. Comparison of human participant results (a) and simulation results (b) for Experiment 2.

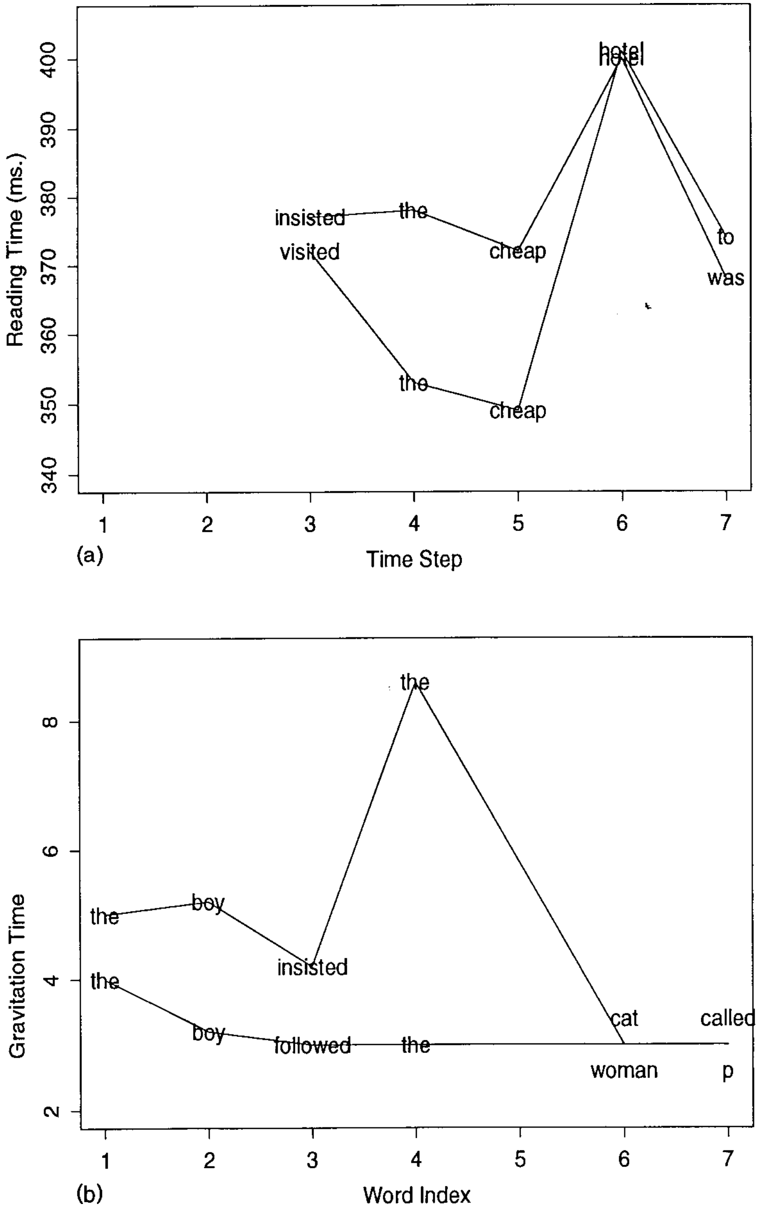


FIG. 9. Reading time at *the* following the verb: pure NP complement verbs versus (nearly) pure sentence complement verbs. Comparison of human participant results (a) and simulation results (b) for Experiment 3.

simulation results near the word *the* following SC-bias and NP-bias verbs (Experiment 3). Again, the human and model results are consistent. This time, both the model and the humans show increased processing time on the first disambiguating word, although, as we noted earlier, in a replication of Experiment 2, the anomaly did not show up until the next word after the disambiguating word (the adjective). The better alignment between the people and the model in these cases may be due to the presence of a stronger signal in the language in each case: the word *that* is very likely to be a complementiser; several of the SC-bias verbs of the experiment cannot occur with direct objects.

The dynamical system for these simulations also has three main attractive regions corresponding to the three major lexical categories: verb, determiner and noun. Again, each region contains several distinct attractors corresponding to the different syntactic uses of words in these classes. Here, the effect of interest shows up on the determiners so we concentrate on the determiner (+ complementiser *that*) subregion to interpret the results.

Figure 10 shows (black dots) the fixed bodies in the determiner (+ *that*) subregion in two dimensions, the first and second principal components of the determiner (+ *that*) sample.[14] In this region, there are three attractors, indicated by the three circles. The attractors correspond to the three grammatically distinct behaviours associated with determiners (and the word *that*) in the training grammar: in the middle right region of the figure, there is an attractor corresponding to subject determiners; in the lower right, one corresponding to object determiners; in the upper left, one corresponding to the complementiser use of *that.* Two distinct clusters lie in the basin of the object determiner attractor: there is a very dense cluster corresponding to unambiguous determiners like *the*, *a*, *one*, etc. ("V[NP]-Det"); there is a more sparsely populated cluster corresponding to the word *that* when it is used as an object determiner ("V[NP]-that"). There are also several clusters which lie in the basin of the subject determiner attractor: there is a dense cluster corresponding to unambiguous subject determiners that occur at the beginning of a sentence ("p-Det"); there are a couple of separate litte clusters corresponding to the word *that* when it is used at the beginning of a sentence ("p-that")—in this grammar, it can only be a subject determiner in this position. There are also two other small clusters in the subject determiner basin, corresponding to determiners following the word *that* and the word *that* following the word *that*—to keep the diagram uncluttered, these are not labelled in Fig. 10. The complementiser basin has only one distinctive cluster in it.

[14] The first three components account for 96% of the variance in the determiner subportion of the visitation set and the first two account for 72 and 20% of the variance, respectively.

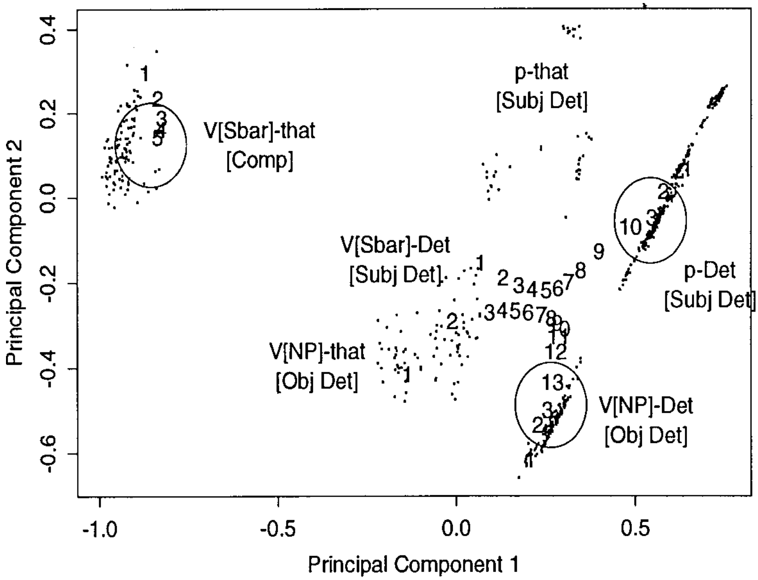


FIG. 10. A portion of the visitation set for simulations 2 and 3 with labelled clusters representing determiners following NP complement and sentential complement verbs.

In all cases, the peripheral clusters in each basin are displaced away from the attractors in a direction that reflects their formal ambiguity: the uses of *that* as a determiner are in the determiner basins, but are displaced in the direction of the complementiser attractor; the uses of unambiguous determiners immediately following SC-bias verbs are (mostly) in the subject determiner basin, but they are displaced simultaneously in the direction of the complementiser attractor and the object determiner attractor, reflecting, respectively, the facts that SC-bias verbs usually take complementisers in this grammar and determiners following verbs usually introduce direct objects.

Note that the clusters "V[Sbar]-Det" and "V[NP]-that" are not very separate from each other in the figure. This is partly a result of the fact that we are only displaying two principal components—these two clusters achieve much better separation on the third component. However, some of the instances of "V[Sbar]-Det" actually lie in the basin of attraction of the object determiner attractor. It took careful tuning of the parameter *p* to make the basin boundary between subject determiner and object determiner land roughly between these two clusters. Note that "misclassification" of some of the "V[Sbar]-Det" instances is not unreasonable—people may sometimes try to parse "the raccoon" in a sentence like "She insists the raccoon . . .” as a direct object. However, this situation may also reflect a technical weakness of the current implementation. It is likely that a revision of the neural network model can make the attractor basins line up perfectly with the grammatically distinct classes (we return to this point in the General Discussion). More generally, however, this result demonstrates that a mix of two behaviours (correct and incorrect “initial” syntactic categorisation) that are distinctly different according to traditional parsers naturally arises out of our system.

Figure 10 makes it apparent how the model makes appropriate reading time predictions in Experiments 2 and 3. An unambiguous determiner like *those* following a transitive verb is processed relatively quickly because the processor starts in the dense cluster near the object determiner attractor and thus has only a small distance to travel to reach the attractor. A trajectory labelled “1 2 3” in the lower right of the figure provides an example of this case. By contrast, the ambiguous word *that* following a transitive verb takes quite a while to process because the processor starts in the “V[NP]-that” cluster, far from the attractor locus (trajectory “1 2 . . . 1 3”). The processing of such a phrase is slowed down both by the influence of the complementiser attractor and by the influence of the subject determiner attractor. This is how the model predicts the major contrast observed in Experiment 2. Likewise, the ambiguous determiner *the* following a transitive verb is processed quickly, just as *those* is (again the trajectory “1 2 3” in the lower right is illustrative). The determiner *the* following an SC-bias verb takes longer to process because it starts far from the subject determiner attractor and is influenced significantly by both the complementiser attractor and the object determiner attractor (trajectory “1 2 . . . 1 0”). In this way, the model predicts the contrast we focused on in Experiment 3. Figure 10 also shows illustrative trajectories for sentence-initial unambiguous determiners (“1 2 3” in the upper right) and for the word *that* following an SC-bias verb (“1 2 3 4 5” in the upper left). These two cases are about equally easy to process, but the “V[Sbar]-that” case has slower trajectories because it is a lower-frequency type but it is isolated enough that it forms its own attractor locus.

Thus these simulations illustrate how competition among attractors provides a natural way of modelling the processing difficulty associated with ambiguous elements. Also, the expected effects of category frequency are observed. Moreover, these simulations make it especially clear how formal similarities can give rise to increased processing time even in circumstances where the system is not making mistakes. We have, in the notion of an attractor basin properly containing an attractor, the tools to encode both major categorical distinctions (by distinguishing basins) and minor, similarity-based influences (by distinguishing elements within a basin in terms of their travel-time to the local attractor).

USING CORPORA TO TRAIN A RECURRENT NETWORK

The models presented in the section headed "Contingent frequencies and *that*", simulated processing of an entire *language*, using samples generated by a finite-state grammar in accordance with frequency patterns modelled on a natural language corpus. These simple global models have advantages over larger-scale models that incorporate a larger number of items. One advantage is that their representations can be analysed more easily. A second advantage is that global models can shed light on how experimental results based on different parts of a language are related to one another. However, there is no guarantee that the properties of these models will generalise to a larger-scale model trained from a real language corpus. Moreover, as we saw in the small-scale model of Experiment 1, there is an important question about which features of natural language to encode in a simulation grammar. Thus using a "toy" grammar to generate a corpus always raises questions about whether the behaviour of the system that depends upon the corpus input is actually due to distortions in the sample corpus. It was therefore important to see whether a model trained from a more realistic subset of a natural language corpus would exhibit the same properties as the models trained from small grammars. This modelling study also provided us with an opportunity to see whether a network trained from a natural corpus could simulate item-specific differences in reading times. This is an important test of the claim that the system is simultaneously sensitive to category-level and to item-specific generalisations.

Following Juliano and Tanenhaus (1994), we trained a recurrent network to predict the complement types that followed past-tense verbs, focusing on those verbs that permit sentence complements to varying degrees. The general architecture of the model is the same as the one used in the previous section (see Fig. 11). However, the simulation presented here differed from that in the previous section along three dimensions: the training set, the input representation and the predictive task the model was given.

The input layer contained 122 units, 72 of which represented verbs; the remaining 50 represented words, word categories and punctuation that could immediately follow the verb. Of the 72 units devoted to each of the verbs, the first 5 units were *similarity* units. Similarity units were assigned values determined from a principal components analysis performed on semantic similarity judgements. The similarity units let the model take advantage of the fact that verbs with similar complement structures typically have related semantics (Fisher, Gleitman, & Gleitman, 1991). While, in principle, a model might learn these similarities from co-occurrence information in a large training corpus, we chose to incorporate similarity into the representation because our training set was sparse and limited in

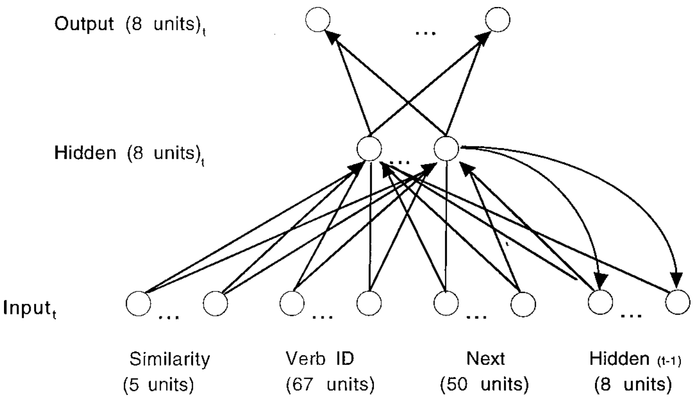


FIG. 11. (Simple) recurrent network from simulation 4. The activation pattern on hidden layer $t - 1$ is set to the activation pattern of hidden layer t from the previous time step (as in Elman, 1990, 1991), indicated by the curved arrow from hidden layer t to hidden layer $t - 1$.

syntactic variability. Models in which we eliminated similarity units showed the same pattern as the model presented here; however, they did not learn as well, and they accounted for less of the variance in human reading times.

The similarity data were generated in a study conducted in collaboration with Joshua Richardson. Three student participants were instructed to judge a set of 70 verbs on how similar in meaning they were to each other on a 9-point scale. Each verb in the set was paired with all of the other verbs, creating 4900 verb comparisons. These were randomly presented to the participants. We averaged the judgements of all three participants and performed a principal components analysis on the 70 × 70 matrix that resulted. The principal components analysis was used to reduce the 70 dimensions to the first five principal components, which together accounted for 60% of the variance. These five components were normalised to create a 5-unit vector for each verb.

The 67 units directly following the similarity units each identified a particular verb. Thus presenting a verb entailed setting the appropriate input values for the five similarity units for that verb and setting its identity unit to 1. The remaining 50 units at the input provided a localist representation for an item that could follow the verb in the corpus. These items could be one of 33 words (frequent articles, pronouns and prepositions), 6 types of punctuation, 10 category types, or “other”, a general category that included the 6% of the items that did not fall into any of the other categories. The full set of verbs and “next” items is presented in Table 6.

The hidden layer and output layer each contained eight units. Each hidden unit was connected to itself and to all of the other hidden units. These

TABLE 6
Verbs and the Words that Followed Them ("Next" Items) from Simulation 4

Verbs			
accepted	advised	asserted	remembered
attempted	bought	bribed	shifted
claimed	conceded	conspired	speculated
crept	crouched	declared	stayed
disputed	drifted	estimated	visited
faded	fared	feared	wrote
felt	flopped	followed	reported
gave	glared	glowed	snarled
grabbed	grew	guessed	sprang
helped	hinted	hoped	tasted
implied	indicated	insisted	vowed
intended	invited	knew	revealed
left	looked	maintained	sounded
mentioned	needed	peered	staggered
pleaded	pledged	predicted	tended
pretended	proposed	protested	wept
realised	recalled	refused	

Words that followed verbs				
!	for	me	RB	those
,	he	my	she	to
.	her	NN	T	up
:	him	NNP	that	us
;	his	NNS	the	VBG
?	in	no	their	VBN
a	it	of	them	we
an	its	on	these	with
at	JJ	other	they	WRB
CD	JJR	our	this	your

recurrent connections among the hidden units are represented in Fig. 11 as $Hidden_{(t-1)}$. In the output layer, each unit represented one of the following complement types: adjectival (ADJ), adverbial (ADV), noun (NP), prepositional (PP), infinitival (INF), sentential (SC), verb (VP) and no complement (NC). The NC category signified end of clause or end of sentence punctuation.

Both the output and hidden units in this model had fixed sigmoid activation functions of the form shown in equation (3):

$$f(\text{net}_i) = \frac{1}{1 + \text{e}^{-\text{net}_i}} \tag{3}$$

where $\text{net}_i = b_i + \sum w_{ij}a_{ij}$, b_i is the bias on unit i, a_i is the activation of unit i, and w_{ij} is the weight from unit j to unit i. The cost function was taken to be the sum of the squared errors at each output unit over all patterns and is shown in equation (4):

$$C = \sum_p \sum_i (a_i - t_i)^2 \tag{4}$$

where i indexes the output units, p indexes the training patterns and t_i is the target value. The error signal was backpropagated through the layers and through one time step to adjust the weights on connections involving hidden units (e.g. Pearlmutter, 1995; Rumelhart et al., 1986).

The training set was taken from the combined Brown corpus and the Wall Street Journal corpus, both tagged and parsed by the Treebank Project (Marcus, Santorini, & Marcinkiewicz, 1993). The training set consisted of all the sentences in the corpus that contained any of our set of 67 past-tense verbs. Thirty of these verbs were targets for our modelling efforts; the other 37 verbs were chosen to create a sample preserving the relative frequency with which different complement types followed past-tense verbs in the full corpus. Table 7 compares the percentage of each of the eight major phrase types found in the Treebank corpus and in the training set. In general, the match is quite good, though sentence complements are somewhat over-represented because all the target verbs permitted sentence complements. Also, verb phrase complements were somewhat under-represented because, in the vast majority of these complements, the past-tense verb was followed by an auxiliary verb. Since individual auxiliary verbs have such high frequencies, the inclusion of even a single auxiliary in the training set would have badly over-represented the frequency of the VP complement category. We decided to overlook this discrepancy between the sample and the corpus because the auxiliary + past-participle pattern did not play a foreseeable role in the phenomena we were studying.

A training instance consisted of a past-tense verb, the word that followed the verb and a complement type. The word pairs and complement types were automatically extracted from the sentences in which they occurred in the corpus. An example is presented in (12).

TABLE 7
Comparison of the Percentage of Major Complement Types Following Past-tense Verbs in the Treebank Corpora (Brown and Wall Street Journal) and Experiment 4 Training Set

	Complement Type							
Source	*ADJ*	*ADV*	*NP*	*PP*	*INF*	*SC*	*VP*	*NC*
Treebank	5.5	5.8	38.7	11.4	5.1	3.4	15.8	7.6
Training set	4.8	4.9	43.1	12.1	6.6	6.3	2.4	6.7

12a. (S (NP (DT The) (NN lawyer)) (VP (VBD insisted) (SBAR (IN that) (S (NP (PRP$ his) (NN client)) (VP (VBD was) (NP (NN innocent)))))

12b. insisted that SBAR

A sentence from the corpus such as the sentence in (12a) would be transformed into (12b). Training the model with an instance from the corpus was carried out in two steps. First, for a verb like *insisted* in (12b), its semantic representation units and its identification unit were activated, then the model was trained to produce the complement type that followed the verb (SBAR in 12b) using backpropagation. Next, the verb input was set to zero and the unit corresponding to the item that followed the verb (*the* in 12b) was activated. The model was again trained to produce the complement. If the specific item that succeeded the verb was not one recognised by the model, then the category of the word was activated. Note also that the interconnections among the hidden units give a sequential component to the model's behaviour. After the first item of a two-item training pair was presented, the hidden unit activation from the current time step was recirculated to the hidden layer in the subsequent time step. Thus in the example above, *insisted* would influence the prediction the model makes when *the* is presented.

The complete training set consisted of patterns from 4798 sentences. The model was initialised with different starting weights for each of three training sessions and trained until the model's overall error (RMS) no longer seemed to decrease. On average, the model made seven passes through the training set.

To generate reading times from the hidden space of this large-scale model, we focused on the behaviour of the model when the word following the verb had been presented (hence including all the cases where a determiner occurred). We used the 4798 sentences of the training set to create a visitation set, which meant that n, the number of fixed bodies in the gravitational model, was 4798. After a little bit of probing, we set $p = 3.0$, $r_{\text{min}} = 0.001$ and $\mu = 0.000005$.

Analysis of the Model

We began with some preliminary analyses of the model to see how the RCN was performing the prediction task. First, we determined that the RCN developed argument structure preferences by examining its predictions for the 10 verbs that most typically occurred with NP complements in the training set, and the 10 verbs that most typically occurred with sentence complements (we later used these same verbs in more detailed comparisons of the model's predictions and human reading time data). After presentation of the verb alone, the RCN correctly assigned the highest likelihood to the complement type that had the highest likelihood in the training set for each of these 20 verbs.

To examine the representations constructed by the RCN, we presented the trained RCN with just the verbs from the training set. We then collected the network's predictions for upcoming complements and performed a principal components analysis on the resulting output activation. Figure 12 shows all the verb types in the space defined by the first two principal components, the latter accounting for 66% of the variance. For purposes of discussion, we will classify verbs according to which complement was assigned the highest probability of the RCN. Different groups of verbs are clearly mapped to specific regions in space, as in the models described earlier. This reflects the fact that items group according to the environment they occur in and the phrase types that they predict. A second observation is that the different groups of verbs vary in how tightly they cluster. In the upper right, there is a tightly knit cluster consisting of verbs that are reliably followed by infinitive complements. Somewhat less tightly knit are verbs that strongly prefer an NP complement. These verbs are for the most part strung out in one dimension along the horizontal axis. Verbs that prefer a

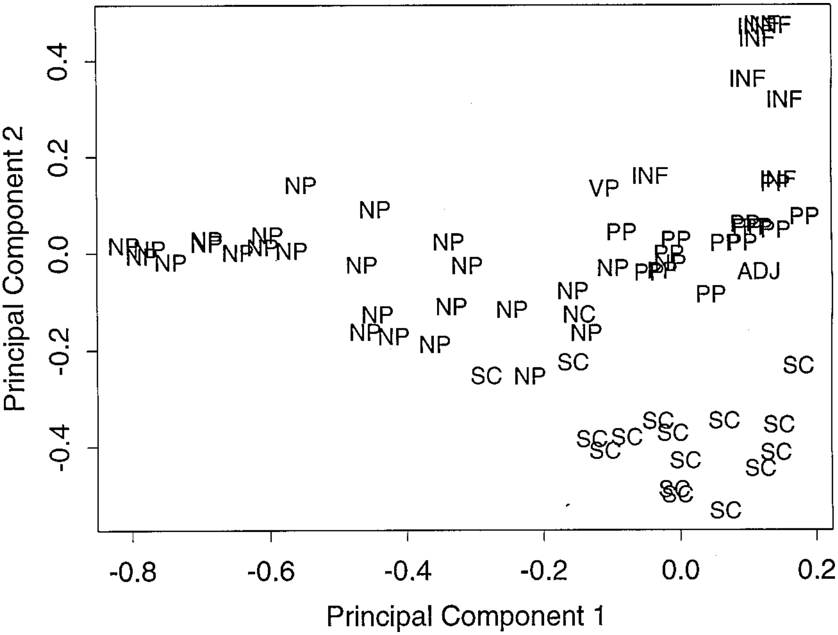


FIG. 12. Analysis (PCA) of predictions from simulation 4 when just the verbs are presented. Each verb is presented by the complement type it predicts in the simulation.

sentence complement form a somewhat loose association, spread along both dimensions. This results from the fact that SC-bias verbs take a variety of complement types. Also note that verbs are not assigned to discrete classes. Rather, they are located on a continuum. Verbs that regularly take both noun phrase complements and sentential complements (as determined by corpus analyses of the Treebank corpus) fill the space between NP-bias and SC-bias verbs, appearing closest to the class that they are most similar to in behaviour.

We also explored whether verb frequency itself had effects above and beyond argument structure preferences. Specifically, to determine whether effects of verb frequency were represented in the model's representational space, we compared the predictions that the model made when presented with each of our experimental verbs with the corpus probabilities of each verb. A plot of both predictions and probabilities in principal components space is shown in Fig. 13. Each arrow represents a verb: the butt of the arrow is its location in corpus probability space, and the head of the arrow represents its position in the model's prediction space. The arrows are longer for lower-frequency verbs compared to higher-frequency verbs [$t(18) = 3.11$, $P < 0.006$], indicating that lower-frequency items are pulled

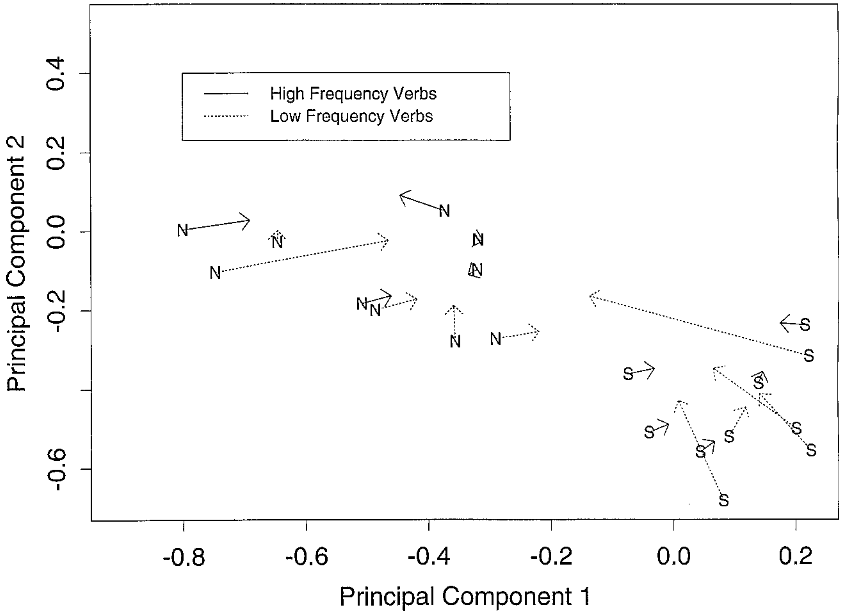


FIG. 13. Attractor effects on items mediated by frequency. N = NP-bias verbs, S = sentential complement verbs.

further from their locations in probability space towards locations associated with other classes.

In the previous section, we presented evidence that dynamical models based on simple grammars with a small number of prototypical lexical items in each class can predict a range of significant qualitative distinctions observable in human reading time data. It is also important to evaluate whether a model based on a more accurate corpus can predict item-specific variation in reading times. To this end, we conducted an experiment using many of the *that* and *that*-less complements that were presented to the large RCN trained on the Wall Street Journal data. We then correlated reading time predictions generated by the gravitational model based on the RCN with the human reading time data.

Experiment 4: Reading Times to NPs in Sentence Complements

Experiment 4 compared reading times to *that* and *that*-less sentential complements when they followed three broad classes of verbs: verbs that are typically followed by a sentential complement (SC-bias verbs); verbs that are typically followed by a noun phrase complement (NP-bias); and verbs that occurred approximately equally with both sentential and noun phrase complements (EQUI-bias). We tried to select verbs in which there was at least an approximate match between the corpus statistics and sentence completion data collected in our lab. It was important to take into account the completion data for two reasons. First, the sample in the Treebank corpus is relatively small and thus subject to errors, especially for lower-frequency verbs. Second, the verbs in the corpus appeared in a variety of syntactic constructions, many of which were quite different from the types of sentences presented in the experiment. The completions provided data for complement preferences in environments similar to those tested in the experiment. We were able to find verbs where the corpus statistics and the completions were in good agreement for the SC-bias and NP-bias verbs, but not for the EQUI-bias verbs, a point we will return to later.

The task we used was the "stop making sense" variant of self-paced reading introduced by Boland, Tanenhaus and Garnsey (1990). We chose this task because it is extremely sensitive to local syntactic and semantic anomalies and the results are more closely time-locked to the input than normal self-paced reading.

Method

Participants. Forty undergraduates from the University of Rochester participated in the experiment for course credit. All were native speakers of English.

Materials. We selected 48 verbs that permitted sentence complements: 16 strongly S-complement biased verbs, 16 strongly NP-complement biased verbs and 16 EQUI-biased verbs. While these classes are useful for grouping the data, the verbs were actually on a continuum with verbs within a class varying in the degree to which they occurred with S-complements. For instance, some of the SC-bias verbs are followed by sentence complements virtually all the time (e.g. *implied*), whereas some are not (e.g. *claimed*). This variability made it especially difficult to select EQUI-bias verbs. Some of the EQUI-bias verbs were closer to verbs in the SC-bias class with regard to the complement types they predicted, while others were more similar to verbs in the NP-bias class. Overall, selection of items for the EQUI-bias class was a compromise—the corpus and sentence completion data were in agreement for only about five of these verbs.

The base frequencies of the verbs were matched within verb class, with each class having similar frequency ranges and distributions. The verbs were embedded in 48 target sentences, each beginning with a determiner-noun NP followed by the main verb, which in turn was followed by either a *that* or a *that*-less sentence complement. The experimental items were combined with 144 filler sentences for a total of 192 trials. The critical sentences were counterbalanced across two lists, such that a verb that appeared with a *that*-less sentence complement in one list appeared with a *that* complement in the other list. Each critical sentence was followed by at least two filler sentences. Roughly 20% of the filler sentences did not make sense at some point for either syntactic or semantic reasons. None of the fillers contained experimental verbs or sentence complement constructions. Overall, 25% of the sentences in each list had a main verb followed by a sentence complement. The target sentences in each list were then pseudo-randomised to create four different orders. This ensured that each item would be rotated through different sections of the list to avoid position effects.

Procedure. Stimuli were presented on a colour monitor attached to an IBM PC with a Digitry CTS timing card. The monitor was set to display 80 characters per line, and all of the critical sentences fit on a single line, with the last word in each sentence accompanied by punctuation. Participants pressed a key on a response box to control the presentation of the sentences. Words accumulated one by one with each key press. Participants were instructed to press the same key as long as the sentence made sense. If the sentence stopped making sense, they were to press a NO button, which terminated the trial. Before the regular experiment began, participants were shown samples of nonsense sentences and received an explanation for each. They then completed 10 practice trials. The entire session lasted about 30 min.

Results

Responses were collected for eight word positions in each sentence beginning with the second word, which was the head noun of the subject. We collected both "stop making sense" judgements (*no*-judgements) and the time between button presses. We will briefly discuss the judgement data followed by the reading time data.

Judgement data. The sum of all the *no*-judgements at each word in the sentence was calculated and then averaged across three regions of the sentence: the verb, the noun phrase that follows the verb, and the succeeding auxiliary phrase. These averages were then converted into percentages. Figure 14 summarises the cumulative percentages of *no* responses for each region. Note that participants rarely judged a sentence to stop making sense until they reached the verb phrase in the complement clause.

The percentage of *no* responses was greater at the disambiguating auxiliary phrase in the complement for *that*-less complements compared to complements with a *that.* However, the degree to which participants rejected the sentences varied markedly with verb class. Participants responded *no* to approximately 20% of the trials in the NP-bias condition, 10% of the EQUI-bias trials and only 5% of the SC-bias trials. At the auxiliary phrase,

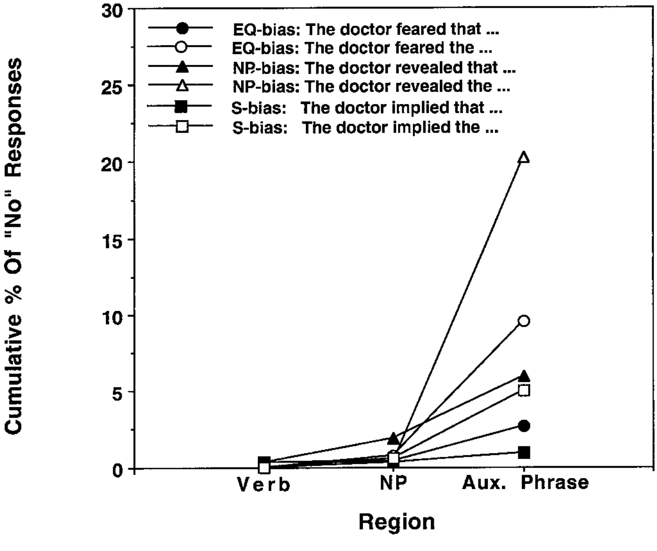


FIG. 14. Cumulative percentage of *no* responses following the different verb types in Experiment 4.

there was a main effect of complementiser on the percentage of *no*-judgements [$F_1(1,38) = 23.66$, $P < 0.001$; $F_2(1,6) = 13.16$, $P < 0.02$] and an interaction between complementiser presence and verb-bias [$F_1(1,72) = 3.70$, $P < 0.05$; $F_2(1,6) = 8.99$, $P < 0.03$]. Simple effects tests showed that the difference between the complementiser present and absent conditions was reliable for the NP-bias verbs [$F_1(1,38) = 39.62$, $P < 0.001$; $F_2(1,6) = 5.34$, $P = 0.06$] and the EQUI-bias verbs [$F_1(1,38) = 6.61$, $P < 0.01$; $F_2(1,6) = 8.99$, $P < 0.03$], but not for the SC-bias verbs.

Reading time data. Recall that reading times to the NP after a *that*-less SC-bias verb should be affected by the strong NP-as-object attractor. Figure 15 presents the difference in reading times at the NP in the *that* and *that*-less conditions. A positive difference reflects longer reading time in the *that*-less condition. The graph shows a large reading time difference for SC-bias verbs, a small difference for the EQUI-bias verbs and no difference for the NP-bias verbs. This was reflected in a verb-type × complementiser interaction [$F_1(2,76) = 4.94$, $P < 0.01$; $F_2(2,12) = 4.56$, $P < 0.05$]. Simple effects showed that the complementiser effect was significant only for the SC-bias verbs [$F_1(1,38) = 35.19$, $P < 0.001$; $F_2(1,6) = 10.05$, $P < 0.02$].

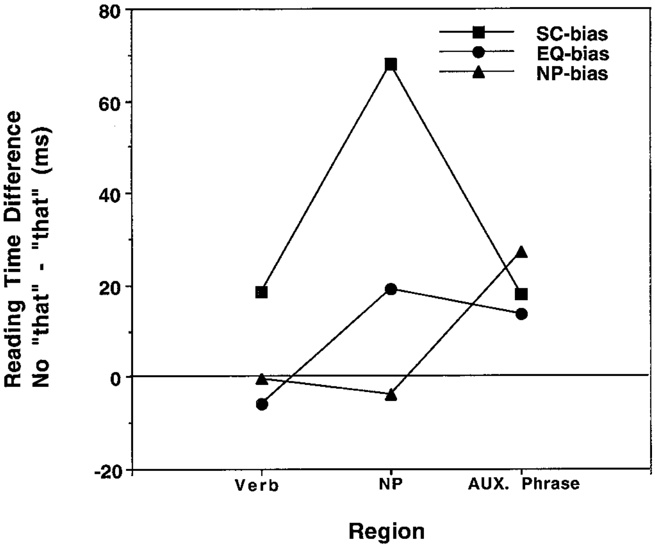


FIG. 15. Reading time differences between the *that* absent and *that* present conditions following different verb types in Experiment 4.

Comparison of the Model's Predicted Reading Times with the Human Data

We measured reading times as the number of time steps to reach an attractor according to equation (2).[15] We then computed a regression in which the model's gravitation time at the word *the* for each target verb was used to predict the difference between human reading times at *the* following a verb and human reading times at *the* following the complementiser *that* which followed the verb. It was important to use the differences to factor out spillover effects because of the frequency difference between "that" and the preceding verb.[16]

The simple regression displayed in Fig. 16 reveals a good positive correlation between the strength of the corpus model's predictions and human reading times at the region following the verb ($r^2 = 0.45$, $P < 0.00005$). It turns out that *the* is treated by the RCN of this simulation as

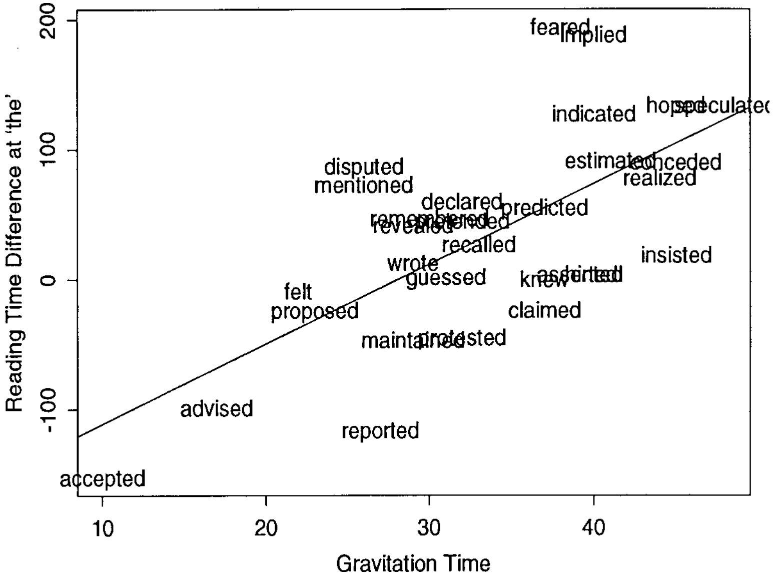


FIG. 16. Regression showing the predicted reading time from simulation 4 as a predictor of human reading difficulty in Experiment 4. $r^2 = 0.45$, $P < 0.00005$.

[15] We selected one of the three simulations at random for this analysis because it is meaningless to average over the hidden unit values of different simulations.

[16] In the model, spillover effects seem to be weaker, so we were able to get away without baselining. Since we did not train the model on sequences of the form "V that the", we had no way to baseline the model in the same way we did the humans.

a very strong indicator of a following NP complement. When any one of these 30 verbs is followed by *the* in the model, the model assigns the highest probability to NP complement. This is, of course, a significant error with respect to the target probabilities for all the SC-bias verbs and about half of the EQUI-bias verbs.[17] Correspondingly, in the gravitational model, all 30 of these cases of verb followed by *the* are in the same attractor basin. The attractor that they are all drawn to corresponds to a prototypical NP-bias verb. The verbs are arranged roughly on a cline according to how frequently they take NP complements versus Sbar complements: NP-bias verbs are closest to the attractor, EQUI-bias verbs are at an intermediate distance, and SC-bias verbs are the farthest away. This cline is correlated with the cline of reading times, so there is a good correlation between gravitation times and reading times.

The Experiment 4 simulations thus provide some indication of the potential of this framework to provide a close quantitative fit to reading time data. Moreover, an interesting further prediction is suggested by the framework. The bias towards predicting an NP complement when a verb is followed by *the* may be too strong in this model. An RCN trained to fit the corpus data more accurately (perhaps by training it longer) would probably give rise to a gravitational model with separate attractors for the SC and NP classes at the word *the*. Such a model would show a non-monotonic relationship between gravitation time and the propensity of the verb to take an NP (*vs* an Sbar) complement: strong NP-bias verbs would have low gravitation times because they are near the NP-attractor; EQUI-bias verbs would have fairly high gravitation times because they are in an intermediate region; while certain very strong SC-bias verbs would have low gravitation times because they would be near the Sbar complement attractor. Several likely instances of very strong SC-bias verbs in this sense (e.g. *said*, *thought*) were not included in the model because their high absolute frequencies would skew the corpus statistics. Nevertheless, the gravitational model leads us to expect that such high-frequency SC-bias verbs will show little reading time difficulty at the word *the*, so we leave this as a prediction to be tested in future research.

GENERAL DISCUSSION

Summary

We proposed a dynamical systems approach to parsing in which syntactic hypotheses are associated with attractors in a metric space and reading times correspond to the time it takes the processing system to gravitate to an

[17] When the verb alone is presented, as we noted above, the model appropriately distinguishes between complement types in most cases.

attractor from where it was initially placed in the space. This system predicts increased reading times when a sequence of words shows mixed evidence as to its proper classification, because mixed evidence results in an intermediate initial placement and hence a relatively long gravitation trajectory. These effects occur even when the word sequence is parsed correctly (i.e. when the processor is initially placed in the basin of attraction corresponding to the correct parse). Thus although the current model treats processing as a process of following a trajectory through a representation space, it is *not* equivalent to serial garden-path models, which assume that slowed reading times following an ambiguous region of a sentence typically reflect an incorrect first parse commitment followed by a time-consuming revision. Rather long reading times often reflect a competition process in the spirit of most constraint-based models. In addition, reading time differences can be associated with differences in the strength of attractors due to lexical frequency effects and the frequency with which higher-order grammatical categories occur in different environments.

Support for the approach came from a series of empirical effects which were predicted by the attractor framework. Experiment 1 showed that syntactic preferences across the same phrase types vary in ways that are predictable based on corpus-derived likelihoods. However, Experiment 2 showed that syntactic preferences are not completely predictable from the grammatical statistics of a corpus—sometimes even ungrammatical possibilities appear to exert an influence (for example, the possibility of treating *that* as a complementiser when it follows a verb that does not permit a sentence complement). Experiments 3 and 4 provided a related example—when reading an NP immediately following a sentence complement verb, readers seemed to be influenced by the hypothesis that the NP was a direct object.

Each of these effects was hypothesised to arise from either attractor competition associated with intermediate representations, or differences in the strengths of attractors associated with frequency differences, or a combination of these. These hypotheses were tested in simulations of Experiments 1–4. The simulations used a recurrent network trained on a word prediction task to (a) construct a similarity-based metric space from a training set generated by a corpus and (b) place the system at a point in that representational space as successive words in an input sentence were presented to the system. We then used an algorithm implementing a gravitational dynamical system to model reading times. More precisely, junctures between words in a corpus were assigned to locations in a representation space according to the similarity of preceding and following word sequences; a sample of such locations was interpreted as a set of point masses; processing involved putting a test body in the space and letting it gravitate towards the mass concentration that drew it most strongly. The model successfully categorised the linguistic input and simulated the reading time effects observed in the experiments. An analysis of the trajectories followed by the dynamical system showed how the reading time contrasts stemmed from the specified principles of dynamical processing.

The Role of the Dynamical Component

Given these initially promising results, it is important to evaluate both the potential contributions of this work and its limitations. The first question that arises is exactly what insights are being contributed by the dynamical component of the model. This question is particularly important because our implementation includes an RCN which performs the essential task of organising instances in the metric space, and one might well wonder what additional information the gravitational component is contributing.

In effect, the dynamical system provides a way of elucidating the organisational system adopted by the RCN and showing how it might plausibly be part of a processing system that generates observed reading time contrasts. In particular, the dynamical system permits us to provide explicit accounts of a number of well-known processing phenomena:

1. The time it takes the system to reach an attractor provides an explicit analog of human reading time.
2. The strength of an attractor corresponds to the enhancing effect of the frequency of a grammatical class on processing time.
3. The possibility of placing elements at contrasting distances from an attractor allows us to conceptually unify two sources of reading time contrasts:
   (a) Word strings containing ambiguous signals are mapped to intermediate clusters which do not form their own attractors and hence take a relatively long time to process.
   (b) Low-frequency elements within a grammatical class are mapped to a large cloud of positions around an attractor and hence take longer to process than the corresponding high-frequency elements, which are mapped to a smaller cloud around the same attractor.[18]

Although we have not emphasised property (b) here, we think it is an important advantage of the dynamical treatment that it allows us to unify increased processing time due to category-based ambiguity and increased processing time due to low frequency, by treating them both as stemming

[18] We have not focused on property (b) here, but it seems generally to be consistent with our implementation, and it is consistent with a number of results relating absolute word frequencies to reading time contrasts (Juliano & Tanenhaus, 1994; Trueswell et al., 1993).

from representational intermediacy. This is well-motivated on information-theoretic grounds: the processor should hedge its bets on categorically ambiguous items because they show a mixture of behaviours; it should also hedge its bets on low-frequency items because it has less experience with them and cannot be as sure of their behaviours.

4. The possibility of having aberrations in the phase space which do not correspond to differences in its topology[19] provides an appealing new approach (though not a full-fledged solution—see below) to the "grain size problem". In essence, it provides a way of thinking about "grain size" in a neural-network compatible medium.
5. The attractors correspond to the behaviourally important syntactic distinctions, and thus embody the "emergent properties" that have sometimes been hypothesised to account for syntactic effects in constraint-based models.

The role of the RCN in the model is to assign the instances of between-word junctures in the test corpus to positions in the representation space according to the statistical similarities among the surrounding-word contexts. In fact, it is fairly apparent in the cases of the simple grammars examined here why the RCN places each element where it does: there are clusters corresponding to junctures between words that are statistically distinct under the training grammar and these are spatially organised according to similarity; the density of elements within a cluster is roughly inversely related to distance from the cluster's centre of mass; junctures that contain locally ambiguous cues show residual effects of the stage when the network was primarily sensitive to immediate context effects—they are on the peripheries of the dense clusters for this reason. These principles are consequences of the well-known sensitivity of neural networks to frequency contrasts and the well-known diffusion of the error signal across layers in backpropagation. With an understanding of these principles, one can, in the case of the simple grammars of Experiments 1, 2 and 3, design a phase space by hand which looks very much like the neural network solutions (which reliably emerge over and over again on repeated training trials with different starting weights) and makes all the desired reading time predictions when the dynamics are applied. Certainly, in more complex grammars, the representational system adopted by an RCN may be harder to intuitively predict. We believe that the type of dynamical analysis proposed here may help identify the important features of the representation in such complex cases.

[19] For example, in the gravitational implementation, there are clusters which do not form their own attractors.

Potential Alternative Approaches

One natural alternative approach to the data we examine here would be to directly associate reading times with some property of the RCN. For example, one might try to use output error itself to predict reading times. While this approach is appealing in some respects, it also has some shortcomings. A model which predicted reading times from output error values would not be explicit about what computation was taking more time in some cases, less time in others. In other words, it would be a merely correlative model. Moreover, such an account would not be very revealing of the representational principles underlying its success: there would be nothing corresponding to a parse hypothesis in the model; there would be nothing revealed about the distinctions between different sources of reading time contrast (e.g. frequency of a class *vs* frequency of an element within a class *vs* multiple class membership of a word). Finally, in several of our models, the network learns "too well" for this approach to work well: certain hidden-layer distinctions which give rise to the appropriate reading time predictions are made very minimal on the output layer because they do not correspond to important behavioural differences in the training environment (e.g. the contrast in reading times between "that" and "those" in "The woman visited that hotel" *vs* "The woman visited those hotels"). Thus it would probably be hard to predict the important contrasts by examining output error alone.

Alternatively, one could model grammatically distinct processing states using a recurrent network by associating them with different points along transient trajectories (as in Rodriguez, 1995; Wiles & Elman, 1995), rather than with fixed points (e.g. attractors). This approach would make direct use of the attractor dynamics of RCNs themselves, rather than creating an additional dynamical overlay as we have done, and such simplicity is certainly desirable. However, it is not obvious to us what features of such models might be likely to correspond to reading times. Nor is it clear to us how to implement even finite-state grammars of any complexity using the transient-based computation principles, which seem to require perceiving certain special kinds of symmetry in the data. Thus while we believe that this approach represents a very interesting direction which may eventually help solve some of the most important problems in dynamical grammar representation, it does not seem useful at the present time for handling the processing data we observe.

Limitations of the Current Account

The current dynamical model, of course, has some serious limitations. Newtonian gravitation is not particularly realistic as a model of neural encoding. Nor is the mixture of RCN mechanisms and gravitational mechanisms particularly elegant. But there are at least two reasons to be interested in such a model anyway: (1) it predicts a range of data in a domain where very few implemented models have been put forth, and (2) Newtonian gravitation of the sort we describe is similar to the settling processes in a variety of neural networks which are somewhat better approximations of brain processing. We see it as an important challenge for future research to develop a neural network implementation that handles a similar range of phenomena. Meanwhile, our account provides a nice stepping stone in this direction because we show not only that a range of data can be implemented in a dynamical model with these kinds of properties, but also why the dynamical model makes the right predictions, and why other models fall short. In this way, we articulate a framework in which to develop a more neurally plausible dynamical model.

One possible implementation, closely related to our current proposal, is to combine a standard RCN hidden layer with a clean-up cycle (e.g. Hinton & Shallice, 1991; Hinton, Plaut, & Shallice, 1993) on the output layer. Such a model would be very similar to the gravitational model we discuss here, with the clean-up cycle doing the work of the gravitation mechanism. A potential technical challenge for this approach stems from the fact that clean-up unit layers tend to form attractors at the corners of the unit hypercube, but in the word prediction task, which we have chosen for its usefulness in studying phrase structure in a statistical setting, the output units need to converge on probability distributions, which are generally far from those corners.

Another possibility is to implement continuous activation settling in the recurrent hidden unit layer, for example by using Pineda's (1995) recurrent backpropagation (RBP) algorithm. RBP forms attractors corresponding to distinct classes defined by the training task, and locates them relative to one another according to principles of formal similarity. One might hope that competition between nearby attractors in this framework would lead to similar results to the ones we have observed here. A challenge for this approach is to handle the long-distance dependencies which are typical in natural language.

The current project is also limited in the range of processing phenomena it addresses. We have focused on reading time contrasts that are specifically related to syntactic class ambiguities. One may wonder if the framework is general enough to handle purely structural ambiguities, for example attachment ambiguities. In fact, structural ambiguities, including attachment ambiguities, are quite generally tied to differences in lexical felicity [e.g. "She watched the policeman with the **binoculars**" (ambiguous) *vs* "She watched the policeman with the **moustache**" (unambiguous)]. These differences are associated with different classes of words and thus give rise, in our model, to distinct attractors associated with distinct structural parses. Thus we expect the model to be able to make appropriate predictions about these kinds of cases as well, and are pursuing this hypothesis in ongoing research.

A Useful Reformulation

Finally, we mentioned above that the current model suggests a new approach to the "grain size" problem. In essence, the introduction of attractors and attractor basins into the metric space representational framework of statistical and neural network language research allows us to specify a "grain" above which behavioural distinctions are grammatically significant and below which they are not. This "grain" is indexed by the parameter p in the gravitational model, which determines which clusters of points will form their own basins of attraction and which will be grouped with other clusters. Thus p is a free parameter.

Nonetheless, the current proposal is an improvement over the processing models based on standard, discrete category grammars discussed by Mitchell et al. (1995), for those models appear to face a "grain paradox" in attempting to model the current data. For example, to treat the reading time slow-down at the word *that* after a transitive verb (Experiment 2) as stemming from a conflation of transitive and SC-bias verbs, such models presumably need to claim that on its first pass, the processor just looks at very broad lexical categories of words, treating all verbs as one type. At the word *that* itself, the principle of minimal attachment chooses the determiner option, since it is compatible with the minimal transitive structure and receives no negative signal, even when lexical details are taken into account, for the verb is transitive and *that* is a legitimate determiner. Choosing a smaller initial grain size that, say, distinguishes transitive from SC-bias verbs, will not help because the verb will select the appropriate structure immediately in this case.[20] In other words, no grain size choice is effective under standard representational assumptions. As researchers begin to document a wider range of statistical effects on parsing, we expect that numerous grain size paradoxes like these will emerge.

In the dynamical model, by contrast, we are able to choose a "grain size" independently of reading time considerations, thus allowing us to make the attractors line up almost perfectly with the behavioural distinctions. With attractors of this grain size, we get the appropriate signal-competition effects in reading times. It cannot be said that the current dynamical treatment solves the grain size problem because we have not explained how the right

[20] Nor would it help to adopt a different principle from minimal attachment, which causes the model to parse the "that N" after a transitive verb as the subject of an embedded sentence. Such a model would incorrectly fail to predict reading time difficulty at the embedded subject noun phrase in sentences of the form "NP V[Sbar] [NP VP_{Sbar}]".

grain size can be selected on the basis of behavioural considerations alone (i.e. independently of the training grammar which language learners clearly have no direct information about). Perhaps it can be said, though, that it transforms what was a grain size paradox into a theoretical question about why a certain grain size might emerge as optimal for learning, representational and processing considerations.

REFERENCES

Abraham, R.H., & Shaw, C.D. (1984). *Dynamics—the geometry of behavior*, Vol. 1. Santa Cruz, CA: Aerial Press.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds), *The cross-linguistic study of sentence processing*. Cambridge: Cambridge University Press.

Bever, T.G. (1970). The cognitive basis for linguistic structures. In J.R. Hayes (Ed.), *Cognition and the development of language*. New York: John Wiley.

Boland, J.E., Tanenhaus, M.K., & Garnsey, S.M. (1990). Evidence for the immediate use of verb control information in sentence processing. *Journal of Memory and Language*, *29*, 413–432.

Brown, P.F., Pietra, V.J.D., DeSouza, P.V., Lai, J.C., & Mercer, R.L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*, 467–479.

Charniak, E. (1993). *Statistical language learning*. Cambridge, MA: MIT Press.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

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

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Ferreira, F., & Henderson, J.M. (1990). The use of verb information in syntactic parsing: A comparison of evidence from eye movements and segment-by-segment self-paced reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*, 555–568.

Fisher, C., Gleitman, H., & Gleitman, L.R. (1991). On the semantic content of subcategorization frames. *Cognitive Psychology*, *23*, 331–392.

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

Frazier, L. (1987). Theories of syntactic processing. In J. Garfield (Ed.), *Modularity in knowledge representation and natural language understanding*. Cambridge, MA: MIT Press.

Frazier, L. (1989). Against lexical generation of syntax. In W.D. Marslen-Wilson (Ed.), *Lexical representation and process*, pp. 505–528. Cambridge, MA: MIT Press.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.

Garnsey, S.M., Pearlmutter, N.J., Meyers, E., MacDonald, M.C., & McConkie, G.W. (1995). The relative contributions of structural biases and plausibility to sentence comprehension. Poster presented at the *36th Annual Meeting of the Psychonomics Society*, Los Angeles, CA, November.

Hanna, J.E., Spivey-Knowlton, M.J., & Tanenhaus, M.K. (1996). Integrating discourse and local constraints in resolving lexical thematic ambiguities. In *Proceedings of the 18th Annual Cognitive Science Conference*, San Diego, CA, July.

Hinton, G.E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Hinton, G.E., Plaut, D.C., & Shallice, T. (1993). Simulating brain damage. *Scientific American*, *269*, 76–82.

Juliano, C., & Tanenhaus, M.K. (1994). A constraint-based lexicalist account of the subject–object attachment preference. *Journal of Psycholinguistic Research*, *23*, 459–471.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.

Just, M.A., Carpenter, P.A., & Woolley, J. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228–238.

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

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

Marcus, M.P., Santorini, B., & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, *19*, 313–330.

Mitchell, D.C. (1989). Verb guidance and other lexical effects in parsing. *Language and Cognitive Processes*, *4*, 123–154.

Mitchell, D.C., Cuetos, F., Corley, M.M.B., & Brysbaert, M. (1995). The linguistic tuning hypothesis: Further corpus and experimental evidence. *Journal of Psycholinguistic Research*, *24*, 469–488.

Pearlmutter, B.A. (1995). Gradient calculations for dynamic recurrent networks: A survey. *IEEE Transactions on Neural Networks*, *6*, 1212–1228.

Perko, L. (1991). *Differential equations and dynamical systems.* New York: Springer-Verlag.

Pineda, F.J. (1995). Recurrent backpropagation networks. In Y. Chauvin & D.E. Rumelhart (Eds), *Backpropagation: Theory, architectures, and applications*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Pritchett, B.L. (1992). *Grammatical competence and parsing performance.* Chicago, IL: University of Chicago Press.

Rodriguez, P. (1995). *Representing the structure of a simple context-free language in a recurrent neural network: A dynamical systems approach.* Technical Report, University of California, San Diego, CA.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland, & the PDP Research Group (Eds), *Parallel distributed processing*, Vol. 1, pp. 318–362. Cambridge, MA: MIT Press.

Schmauder, A.R., & Egan, M.C. (1995). Verb subcategorization information, argument fit, and on-line sentence processing. Poster presented at the *36th Annual Meeting of the Psychonomics Society*, Los Angeles, CA, November.

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

Spivey-Knowlton, M., & Tanenhaus, M.K. (submitted). Syntactic ambiguity resolution in context: Modeling the effects of referential context and lexical frequency within an integration-competition framework.

Spivey-Knowlton, M., Trueswell, J.C., & Tanenhaus, M.K. (1993). Context effects in syntactic ambiguity resolution. *Canadian Journal of Experimental Psychology*, *47*, 276–309.

Strogatz, S. (1994). *Nonlinear dynamics and chaos.* Reading, MA: Addison-Wesley.

Tabor, W. (1994). The gradual development of degree modifier "sort of" and "kind of": A corpus proximity model. In *Proceedings of the 29th Annual Meeting of the Chicago Linguistics Society*, Chicago, IL, April.

Tabor, W. (1995). Lexical change as nonlinear interpolation. In *Proceedings of the 17th Annual Cognitive Science Conference*, Pittsburgh, PA, July.

Tanenhaus, M.K., & Trueswell, J.C. (1995). Sentence comprehension. In J. Miller & P. Eimas (Eds), *Speech, language, and communication*, Vol. 11, pp. 217–262. San Diego, CA: Academic Press.

Trueswell, J.C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, *35*, 566–585.

Trueswell, J.C., Tanenhaus, M.K., & Kello, C. (1993). Verb-specific constraints on sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 528–553.

Trueswell, J.C., Tanenhaus, M.K., & Garnsey, S.M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic disambiguation. *Journal of Memory and Language*, *33*, 285–318.

Wiles, J., & Elman, J.L. (1995). Landscapes in recurrent networks. In *Proceedings of the 17th Annual Cognitive Science Conference*, Pittsburgh, PA, July.

Zipf, G.K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Reading, MA: Addison-Wesley.