The Information Conveyed by Words in Sentences

John Hale[1]

A method is presented for calculating the amount of information conveyed to a hearer by a speaker emitting a sentence generated by a probabilistic grammar known to both parties. The method applies the work of Grenander (1967) to the intermediate states of a top-down parser. This allows the uncertainty about structural ambiguity to be calculated at each point in a sentence. Subtracting these values at successive points gives the information conveyed by a word in a sentence. Word-by-word information conveyed is calculated for several small probabilistic grammars, and it is suggested that the number of bits conveyed per word is a determinant of reading times and other measures of cognitive load.

KEY WORDS: computational psycholinguistics; entropy reduction.

INTRODUCTION

Ambiguity resolution is perhaps the central problem (Tabor & Tanenhaus, 2001) in sentence processing. How is it that human sentence understanders are able to recognize combinatory relationships, from an infinite range of possibilities, to arrive at a meaningful interpretation of a spoken or written sentence?

This question is typically addressed experimentally, using some measure of cognitive load to reason backward to the choices a sentence understander has made in on-line processing. The present paper inverts this usual arrangement by showing how a fairly general conception of ambiguity resolution is sufficient to characterize a range of cognitive load patterns. Ambiguity resolution in this sense is the elimination of impossible structural

101

analyses for a string—a natural measure of the cognitive "work" a comprehender does on the way to determining the speaker's intended analysis.

This demonstration implies that some aspects of cognitive load as measured in reading time experiments follow simply from the statement of the sentence understanding problem as an ambiguity resolution problem. It also suggests a way that explicit linguistic knowledge—both symbolic combinatory knowledge and numerical knowledge expressed as probabilities—can be used by comprehenders. In doing so, it adopts the competence hypothesis (Chomsky, 1965) that our knowledge of language is directly used in comprehension.

The main claim is that cognitive load is related, perhaps linearly, to the reduction in the perceiver's uncertainty about what the producer meant. Historical antecedents and a few closely related proposals are examined in a brief section immediately following this introduction. The central section formalizes the main claim by appealing to information theory, in a way that avoids the criticisms of Chomsky (1956), by using a phrase-structured language model. The penultimate section shows how the claim derives processing predictions about constructions that have been well studied in the sentence processing literature. The final section concludes with some speculations about connectionist networks whose operation accord with the main claim.

THE TERRAIN

Psycholinguistic research on human syntactic processing has always sought to integrate a wide variety of findings into explicit, general theories. This kind of synthesis, toward which the current paper strives, is often very difficult because of the range of natural language phenomena a processing theory must confront.

For instance, in empirical work over the past two decades, evidence has been accruing that the human sentence processor is sensitive to gradient factors such as frequency (MacDonald, Pearlmutter, & Seidenberg, 1994; Mitchell, Cuetos, Corley, & Brysbaert, 1995; Trueswell, 1996), thematic fit, (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Trueswell, Tanenhaus, & Garnsey, 1994), and pragmatic felicity (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). These findings are reviewed by Tanenhaus and Trueswell (1995) and Gibson and Pearlmutter (1998). At the same time, there has been a corresponding revival of interest on the theoretical side in probabilistic grammars as explanations for these effects, as well as core phenomena not previously viewed as gradient.

Although strongly motivated by recent findings, the idea of using probabilistic grammars in psycholinguistic research is an old one, going back to

the work of Patrick Suppes. Suppes (1970) proposes the specific formulation of probabilistic grammar used in this paper as an account of the child's developing knowledge of language. He points out that because probabilistic grammars define the frequencies of words and phrases in a language, they are like any other statistical model whose fit to a given sample can be evaluated in a standardized way. This view, adopted in the present work, takes probabilities to be linguistic properties that are predicted by a grammar, on a par with acceptability distinctions.

Explicit mechanisms of probabilistic *processing* have been pursued most intently by connectionists. In the PDP tradition (McClelland & Kawamoto, 1986; McClelland & St. John, 1989; St. John & McClelland, 1990) it has long been the norm to view constraints on sentence processing, if not grammar itself (Legendre, Miyata, & Smolensky, 1990a, 1990b) as being partial and numerically valued. This view takes on a new form in more recent work (Elman, 1990; Rohde, 2002; Tabor, Juliano, & Tanenhaus, 1997) in which back-propagation is applied to learning the string sets and frequencies of given grammars. A drawback to this approach is that the internal states of the induced machines are often uninterpretable (Steedman, 1999). This criticism takes on a diminished relevance as the application of dynamical systems theory to these states becomes more and more refined (Rodriguez, 1999; Tabor, 2000).

At issue more centrally are the kinds of general principles that (perhaps imperfectly) characterize the operation of sentence processing models estimated from linguistic data.

For example, on the Visitation Set Gravitation model of Tabor *et al.* (1997) reading times can be derived from a post-hoc analysis of a trained Simple Recurrent Network (SRN). This analysis yields a landscape of attractors from the records of hidden unit activations. By observing how long it takes a particular hidden unit state (representing a word along with its left-context) to "gravitate" into an attractor (possibly representing a kind of semantic integration), Tabor *et al.* obtain a measure of the work a comprehender does integrating a word into a developing analysis.

In these experiments, it can take longer to settle on an attractor when many competing attractors—parser states—need to be considered. This is a joint consequence of the gravitational parameters, and the localization, by back-propagation, of similar parser states near one another. It may be that high predicted reading times are the result of the parser traversing "confusing" regions of hidden unit space where the influence of many attractors is simultaneously felt. This supposition motivates the development, in the next section, of a notion of uncertainty or confusion that might characterize, at a high level, what the Visitation Set Gravitation model is doing.

An independent line of work (Jurafsky, 1996; Narayanan & Jurafsky, 1998, 2002) with Bayesian nets (Jensen, 1996; Pearl, 1988) takes a more strongly grammatical approach in the tradition of localist connectionism (Feldman & Ballard, 1982). A Bayesian net is a kind of graphical model that can be hand-crafted to represent a distribution over linguistic variables, such as constructions or lexical entries. Each node of the net is associated with a conditional probability table expressing a distribution on values of the node's variable. For instance, a constituent like NP (noun phrase) might be associated with a distribution over {true, false} encoding the net's belief that the constituent is really present or not. The edges between nodes have a conditional probability interpretation, so they can be estimated directly from large corpora. Evidence is inserted into a Bayesian net by setting the state of "observable" nodes and then propagating probabilities across the edges of the graph according to Bayes' Law.

Jurafsky and Narayanan examine how the probabilities change as the net's beliefs are updated in light of new evidence. The approach is strongly grammatical because the node probabilities can be immediately interpreted in terms of a linguistic analysis. For instance, Jurafsky (1996) discusses the simultaneous consideration of two constructions, the MAIN-CLAUSE-NON-SUBJECT-WH-QUESTION and the SUBJECT-MATRIX-WH-QUESTION upon being presented with the initial string "who can. . . ." If one construction becomes overwhelmingly plausible, as a result of any combination of evidence, Bayesian belief propagation allows it to "explain away" evidence for the alternatives. The probabilities of individual constructions represented in the Bayes net can go up and down as processing proceeds. Using the idea that the parser pursues only the highest probability analyses (Gibson, 1991; Kurtzman, 1985), the Bayesian approach predicts garden path effects when an analysis falls below some threshold and is forgotten, but is later required when no other alternative is viable.

Jurafsky and Narayanan's predictions about human processing derive from probabilities of representations defined by a grammar, so their work respects the competence hypothesis mentioned in the introduction. However, these calculations are carried out over Bayesian nets for complete linguistic analyses. It is not clear if a dynamically constructed Bayes net (Charniak & Goldman, 1993) would derive the same predictions. At the same time, such partial networks would seem to be called for to cover the full set of constructions defined by a recursive grammar. This kind of summation over an infinite number of linguistic representations is implicitly performed by using a closed-form solution of an infinite series (see equation 6) in the next section to calculate psycholinguistic predictions about cognitive load.

Finally, Den and Inoue (Den & Inoue, 1997; Inoue & Den, 1999) follow Jurafsky in endorsing a beam-search interpretation of garden pathing.

In this pioneering application of information theory to sentence processing, the analyses considered by a parallel parser are ranked according to the entropy of the distribution of verbs licensed to occur in a particular syntactic configuration. Den and Inoue address the puzzle of comprehenders' expectations for particular verbs following sequences of candidate arguments in verb-final languages such as Japanese. Their Verb Predictability Model would seem to be a special case of the more general architecture to be proposed in the following section, which is sensitive to the predictability of all structures, not just verbs. However, this proposal differs with Den and Inoue in avoiding any appeal to beam search or reanalysis.

SENTENCE PROCESSING AS ENTROPY REDUCTION

This section shows how to derive processing predictions about reading time from probabilistic grammars encoding linguistic generalizations that are both categorial and numerical. Three simplifying assumptions place general constraints on sentence processing as an ambiguity resolution problem.

1. During comprehension, sentence understanders determine a syntactic structure for the perceived signal.
2. Producer and comprehender share the same grammar.
3. Comprehension is eager; no processing is deferred beyond the first point at which it could happen.

These assumptions suppose that there are combinatory relationships among words presented during incremental sentence processing. A probabilistic grammar is known to both speaker and hearer; the derivations of this grammar completely determine the combinatory relationships to be recognized. Because of ambiguity, however, there is uncertainty about which derivation the speaker intends. This uncertainty is especially great when only an initial segment of a sentence has so far been presented. But if the processor performs disambiguating work, this uncertainty can be expected to go down as more words are presented.

If uncertainty about a derivation measures the total amount of ambiguity-resolution work a processor will have to do, then reduction in this uncertainty should measure the maximal amount of work that can be done between one word and the next. This is the amount of work done by an *eager* sentence processor.

The reduction in uncertainty from one word to the next is the information conveyed by that word. To the extent that sentence comprehension is eager, this information conveyed should closely match other word-by-word measures of information processing.

PROBABILISTIC CONTEXT-FREE GRAMMARS

I will adopt the formalism of probabilistic (or stochastic) context-free grammars (PCFGs) to make the discussion of derivations on a grammar explicit. More complete presentations can be found in all modern computational linguistics texts (Charniak, 1993; Jurafsky & Martin, 2000; Manning & Schütze, 2000). Intuitively, a PCFG is just the kind of phrase structure grammar familiar from linguistics, augmented with probabilities on the rules. The rules of an example grammar are given in Fig. 1.

The rules in Fig. 1 have the property that the probabilities of all rules expanding the same nonterminal symbol sum to 1.0. Grammars with this property are "normalized", which, by itself, is not enough to guarantee a proper probability model (see Appendix A). Figure 1 also shows how the idea of derivational ambiguity carries over from formal language theory. These rules generate the string "the cop saw the spy with binoculars" by two different derivations, given in Fig. 2.

Note first that all derivations begin with the start symbol S. The derivation on the right side of Fig. 2 has probability $(0.4)^3 = 0.064$. The derivation on the left side of Fig. 2 has probability $(0.4)^2 \times 0.6 \times 0.2 = 0.0192$. The probability of a generated string is just the sum of the probabilities of all derivations that generate it: $0.0192 + 0.064 = 0.0832$. The idea of sentence processing as entropy reduction is that all of the work that the processor does is like the kind of ambiguity resolution needed to decide between these two derivations, and that the magnitude of this work is measured by reading time.

Entropy

Ambiguity-resolution work can now be formalized as the information conveyed by a word in a sentence generated by a PCFG. Essential to the definition of information conveyed is the notion of entropy (Shannon, 1948). The entropy of a random variable is the uncertainty, or missing

$$
\begin{array}{llll}
1.0 & S   & \rightarrow & NP\ VP \\
0.4 & NP & \rightarrow & \text{the spy} \\
0.4 & NP & \rightarrow & \text{the cop} \\
0.2 & NP & \rightarrow & NP\ PP \\
1.0 & PP & \rightarrow & \text{with binoculars} \\
0.6 & VP & \rightarrow & \text{saw } NP \\
0.4 & VP & \rightarrow & \text{saw } NP\ PP \\
\end{array}
$$

Fig. 1. Example probabilistic context-free phrase structure grammar.

```
S                                      S
NP VP                                  NP VP
the cop VP                             the cop VP
the cop saw NP                         the cop saw NP PP
the cop saw NP PP                      the cop saw the spy PP
the cop saw the spy PP                 the cop saw the spy with binoculars
the cop saw the spy with binoculars
```

Fig. 2. Two derivations of "The cop saw the spy with binoculars."

information associated with that random variable. More explicitly, for a discrete random variable X with outcomes $x_1, x_2, \ldots$ having probabilities $p_{x_1}, p_{x_2}, \ldots$ the entropy $H(X)$ is

$$H(X) = -\sum_{x \in X} p_x \log_2 p_x \tag{1}$$

The form of equation 1 shows that entropy is identical with the expected surprisal $\log_2 p_x$ of an unknown outcome. Different distributions on the p_{x_i} lead to different entropies, measured in bits when the base of the logarithm is 2. When all outcomes are equally likely, entropy is maximal. For example, the entropy of a fair die is about two and a half bits (Fig. 3).

Under the assumption that sentence understanders determine a syntactic structure for the sentences being understood, the random variable of interest must be one whose outcomes completely determine syntactic structure. The set of derivations on a PCFG fits this bill. A potentially infinite but nonetheless discrete set, they encapsulate everything there is to know about the syntactic structures the PCFG defines. Denoting the grammar at

$$DIE = \begin{array}{c|c} \text{value} & \text{probability} \\ \hline 1 & \frac{1}{6} \\ 2 & \frac{1}{6} \\ 3 & \frac{1}{6} \\ 4 & \frac{1}{6} \\ 5 & \frac{1}{6} \\ 6 & \frac{1}{6} \end{array} \qquad p_{(DIE=3)} = \frac{1}{6}$$

$$\begin{aligned} H(DIE) &= -\sum_{x \in DIE} p_x \log_2 p_x \\ &\approx 2.58 \text{bits} \end{aligned}$$

Fig. 3. Entropy of a fair die.

hand by *G,* let TREE$_G$ be a random variable whose values are derivations on *G,* and let *W* be a string-valued random variable whose outcomes are in the language of *G.* Then the information conveyed by the first *i* words of a sentence generated by *G* is

$$I(\text{TREE}_G \mid W_{0\cdots i}) = H(\text{TREE}_G) - H(\text{TREE}_G \mid W_{0\cdots i}) \qquad (2)$$

As standardly defined (Cover & Thomas, 1991), information conveyed is the reduction in entropy of one random variable by discovering the outcome of another. To characterize the information conveyed to the human sentence processor, consider the information *I* conveyed by just the *i*th word w_i, given the preceding words. Equation 3 expresses this quantity as a difference in the conditional entropy of TREE$_G$.

$$I(\text{TREE}_G \mid W_{0\cdots i}) - I(\text{TREE}_G \mid W_{0\cdots i-1}) = [H(\text{TREE}_G) - H(\text{TREE}_G \mid W_{0\cdots i})]$$

$$- [H(\text{TREE}_G) - H(\text{TREE}_G \mid W_{0\cdots i-1})]$$

$$I(\text{TREE}_G \mid W_i = w_i) = -H(\text{TREE}_G \mid w_{0\cdots i}) + H(\text{TREE}_G \mid w_{0\cdots i-1})$$

$$= H(\text{TREE}_G \mid w_{0\cdots i-1}) - H(\text{TREE}_G \mid w_{0\cdots i}) \quad (3)$$

Equation 3 says that the answer to the question "how much information does a comprehender get from a word?" is the same as the answer to "how much was uncertainty about the derivation reduced?" The answer to this latter question is found in the work of Ulf Grenander (1967), which shows how to calculate the entropy of *all* derivations rooted in a particular nonterminal symbol. Grenander's theorem 4.2 expresses the entropy of a grammar symbol as the sum of two quantities:

1. the entropy of the single-rule rewrite decision and
2. the expected entropy of any children.

These are the two terms of equation 5 whose interpretation is now considered more closely.

Continuing to symbolize as *G* some given PCFG, let the set of production rules in *G* be $\prod$. For a given nonterminal ξ the finite set of rules rewriting ξ is denoted $\prod(\xi)$. Define lowercase *h* to be a vector indexed by nonterminal symbols. Each component given in equation 4 contains the entropy of a single rewrite decision, whose outcomes are rule choices *r* $\Pi(\xi)$ having probability p_r.

$$h_i = h(\xi_i) = -\sum_{r \in \Pi(\xi_i)} p_r \log_2 p_r \qquad (4)$$

Grenander found a recursion relation for the entropy of nonterminals in terms of *h* and the expected entropy of the resulting daughters. Equation 5

depicts[2] the general situation in which rule r rewrites a nonterminal ξ_i as n daughters, $\xi_{j1}, \xi_{j2}, \ldots, \xi_{jn}$.

$$H(\xi_i) = h(\xi_i) + \sum_{r\in\Pi(\xi_i)} p_r\left[H(\xi_{j1}) + H(\xi_{j2}) + \cdots\right] \tag{5}$$

Letting A be the expectation matrix of grammar G (see Appendix A), the solution to this recursion can be expressed succinctly as the matrix equation 6 where I is now the identity matrix and h is given component-wise by equation 4.

$$H_G = (I - A)^{-1}h \tag{6}$$

Consistency ensures that $(I - A)^{-1}$ exists. H_G is then a vector indexed by the nonterminals of G. Because all derivations begin with the start symbols S, $H_{G(S)} = H(\text{TREE}_G)$.

Conditional Entropy

Grenander's result shows how to compute $H_{G(X)}$ for all nonterminals X in one step by inverting a matrix. However, to calculate how much information a comprehender gets from a word using equation 3, the conditional entropy $H(\text{TREE}_G|w_{0\cdot\cdot i})$ is needed. Grenander's recursion relation is again applicable, but only to derivations resulting in the left-prefix $w_{0\cdot\cdot i}$. Care must be taken that the entropy of common subderivations in this set only be counted once.

The algorithm in Fig. 4 is designed to do this counting. Given a set D_A of derivation trees all rooted in the same nonterminal, it recursively computes the entropy of the tree set compatible with the prefix seen so far by applying equation 5. H_{expected} for a string is simply the sum of the entropies for each symbol on the grammar as specified by equation 6. The contribution of this paper is the observation that intermediate states of a top-down parser are specifications of the classes of derivations that can derive the left-context, followed by any continuation. This point has been made and applied quite productively by Lang and colleagues (Billot & Lang, 1989; Lang, 1974). What has not been observed is that the same methods for calculating the entropy of a nonterminal can be applied to the grammars implicitly defined by intermediate parser states. This permits the straightforward definition of "information conveyed by a word in a sentence," which, I suggest, is a quantity of some psycholinguistic interest.

---

[2] N. B. the definition of uppercase H, the entropy of a nonterminal, uses lowercase h, the entropy of the single-rewrite decision.

procedure $H(\mathcal{D}_A)$

  initialize the result $temp = 0$
  partition $\mathcal{D}_A$ into k classes sharing the same rule $A \to \gamma$ weighted with probability p
  divide the probabilities p by their sum to obtain k renormalized probabilities p_{norm}
  for $i = 1$ **to** k **do**
    add $p_{\text{norm}_i} \log_2 p_{\text{norm}_i}$ to $temp$
    let n be the minimum number of expanded daughters of A across derivation trees in class i
    let β be the substring of γ from the nth symbol to the end
    let v be a vector indexed from 1 to n
    for $j = 1$ **to** n **do**
      set v_j to be the union of all derivation trees rooted in the jth nonterminal from class i
      add $p_{\text{norm}} [H(v_1) + H(v_2) + \cdots + H(v_n) + H_{\text{expected}}(\beta)]$ to $temp$ *(* recur on big H *)*
    end for
  end for
  return $temp$

Fig. 4. Computing the entropy of the partial derivations common to D_A.

Left Recursion

There is one rather technical barrier to applying this idea. Even leaving some nonterminals unexpanded, the set of partial derivations generating some prefix of a sentence may not be finite. This is the problem faced by top-down parsers on left-recursive grammars: there is an infinity of possible analyses indexed by the number of cycles through applicable, left-recursive rule sets. A finite factorization of this infinity is needed so that the entropy of the distribution on derivations can be calculated.

Rather than modifying the top-down parser, this problem can be dealt with by transforming the grammar to remove left recursion (Huang & Fu, 1971). Claims of strong equivalence are then justified to the extent that the transform is invertible.

EXAMPLES

Main-Verb/Reduced Relative

The method described above provides a characterization of the famous garden path effect of Bever (1970) in terms of information conveyed. It suffices to consider only syntactic processing, because semantic rules are taken to be in one-to-one correspondence with syntactic rules (Steedman, 2000). Following Crain and Fodor (1985) and Gazdar, Klein, Pullum, and Sag (1985), no further processor resources beyond those needed for parsing probabilistic context-free phrase structure grammar are assumed.

```
1.00   S        →   NP VP
0.88   NP       →   DT NN
0.12   NP       →   NP RRC
1.00   PP       →   IN NP
1.00   RRC      →   Vppart PP
0.50   VP       →   Vpast
0.50   VP       →   Vppart PP
1.00   DT       →   the
0.50   NN       →   horse
0.50   NN       →   barn
0.50   Vppart   →   groomed
0.50   Vppart   →   raced
0.50   Vpast    →   raced
0.50   Vpast    →   fell
1.00   IN       →   past
```

Fig. 5. Initial Bever grammar.

Grammar

Consider the grammar of reduced relative clauses shown in Fig. 5. Key probabilities of this Initial Bever Grammar are set according to corpus frequencies compiled by Rohde (2002).

The morphology of "raced" is ambiguous between the past participle and the simple past tense. Although lexical in origin, this ambiguity causes constructional ambiguity because it permits several possible phrase structures for the sentence "the horse raced past the barn."

The grammar in Fig. 5 expresses the traditional adjunction analysis of relative clauses with the left-recursive rule NP → NP RRC. This left recursion is removed in the grammar of Fig. 6. In this weakly equivalent but

```
1.00   S        →   NP VP
1.00   PP       →   IN NP
1.00   RRC      →   Vppart PP
0.50   VP       →   Vpast
0.50   VP       →   Vppart PP
1.00   DT       →   the
0.50   NN       →   horse
0.50   NN       →   barn
0.50   Vppart   →   groomed
0.50   Vppart   →   raced
0.50   Vpast    →   raced
0.50   Vpast    →   fell
1.00   IN       →   past
0.88   NP       →   DT NN
0.12   NP       →   DT NN Z0
0.88   Z0       →   RRC
0.12   Z0       →   RRC Z0
```

Fig. 6. Non–left-recursive Bever grammar.

nonterminal	entropy
S	5.07
PP	2.05
RRC	3.05
VP	3.02
DT	0.00
NN	1.00
Vppart	1.00
Vpast	1.00
IN	0.00
NP	2.05
Z0	4.09

Fig. 7. Entropies of nonterminals in non–left-recursive Bever grammar.

non–left-recursive grammar, the adjunction analysis is recoded as comple-
mentation in a way that can be seen in Fig. 10, to be discussed later.

Applying equation 6, the entropy of each nonterminal is given in Fig. 7.

States of a Top-Down Parser

The parser begins analyzing "the horse raced past the barn fell" having
heard no words at all. At this point, the conditional entropy of the parser
state given the (nonexistent) input string is just the entropy of the grammar
$H_{G(S)} = 5.07$ bits.

Upon hearing the first word "the," two classes of analyses come into
play, corresponding to the two ways the noun phrase could have been
expanded. In Fig. 8 and subsequent depictions, nonterminals with asterisked
names are unexpanded.

Considering only these classes of analyses, the conditional entropy of the
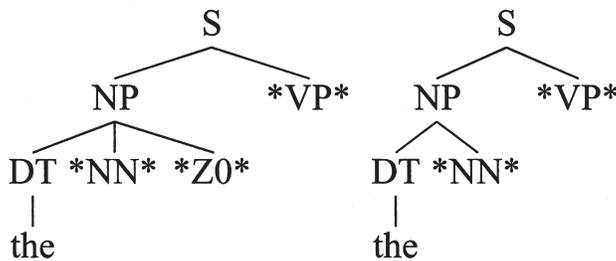start symbol S turns out to be equal to its unconditional entropy—because

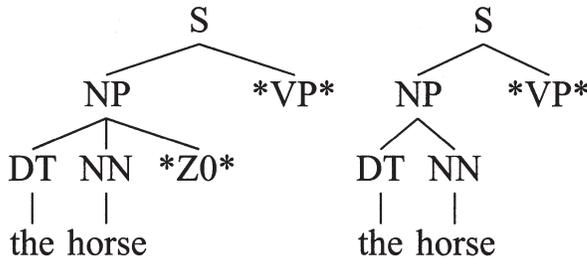**Fig. 8.** Parser state having heard "the."

Fig. 9. Parser state having heard "the horse."

every sentence on this grammar starts with "the," the word conveys no information at all. However, the next word, "horse," is informative. As shown in Fig. 9 the word "horse" following "the" conveys one bit because the speaker has chosen one of two equally probable alternative ways of expanding NN.

At "the horse raced" the parser state comprises four analyses, shown in Fig. 10.

At the word "raced," the parser first explicitly represents the distinction between main verb and reduced relative structures.
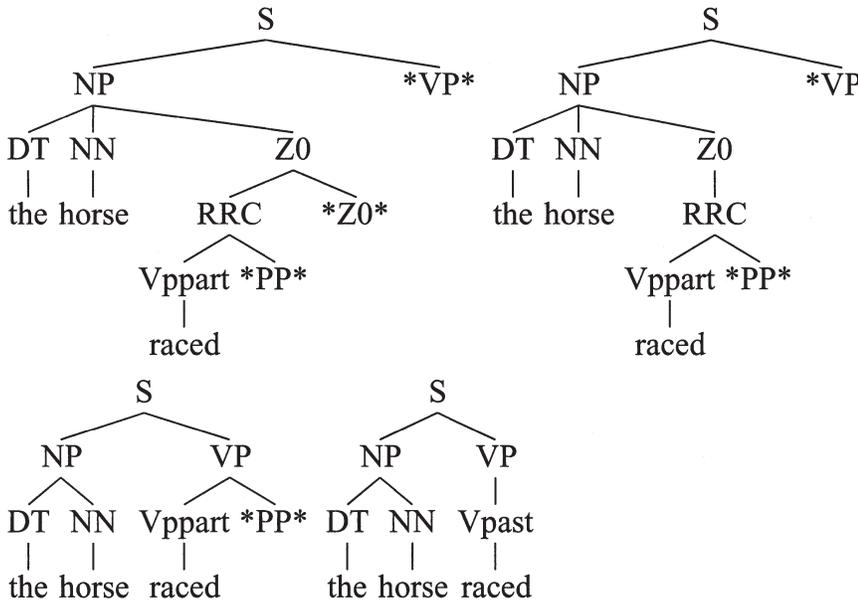

Fig. 10. Parser state having heard "the horse raced."

The rest of the words are processed similarly. For words that reduce entropy, the magnitude of the reduction is written in Fig. 11. At "fell," where the number of compatible derivations is reduced to one (from six), nearly four bits are conveyed—approximately 75% of the information specified by the grammar. This is the sense in which the garden path effect is characterized as confusion brought on by the sheer volume of information being processed. In rejecting the distinction between most highly valued, within-beam, and out-of-beam analyses (Frazier & Clifton, 1996), this account requires relatively few assumptions compared to alternatives that invoke particular structure-building operations, beam-width constants, or semantic contexts.

The NP/S and NP/Z Ambiguities

The sort of predictions afforded by an information-based account are numerical; thus they can also be used to characterize processing asymmetries *between* sentences as well as *within* sentences.

The NP/S and NP/Z temporary ambiguities, illustrated below in (1) are asymmetric in just this way (see also Gibson, Argaman, & Babyonyshev, this volume).

(1) a. The Australian woman saw the famous doctor had been drinking quite a lot.
    b. Before the woman visited the famous doctor had been drinking quite a lot.

The NP/S ambiguity comes about because the verb "saw" can either take an SBAR or an NP complement; the human sentence processing mechanism is lured up the garden path by the NP complement analysis until it reaches the auxiliary "had," where it becomes apparent that only the SBAR subcategorization can be correct. Likewise, the NP/Z ambiguity is due to the verb "visit" taking either zero or one NP complements.

Sturt, Pickering, and Crocker (1999) found that the garden path effect for the NP/Z ambiguity (1b) was more than four and a half times the size

word	reduction in entropy (bits)
the	0
horse	1
raced	0.123
past	0
the	0
barn	0.123
fell	3.82

Fig. 11. Bits of entropy reduction for a garden path sentence.

0.75	S	→	NP VP
0.25	S	→	PP SBAR
1.00	SBAR	→	NP VP
1.00	NP	→	SPECNP NBAR
1.00	SPECNP	→	DT
1.00	NBAR	→	N
0.25	VP	→	V[SUBCAT2] NP
0.25	VP	→	V[SUBCAT1]
0.25	VP	→	V[SUBCAT5] SBAR
0.25	VP	→	V[SUBCAT4,ASP] VBAR[PRP,COP]
1.00	VBAR[PRP,COP]	→	V[SUBCAT7,PRP,COP] VBAR[PRP]
1.00	VBAR[PRP]	→	V[SUBCAT2,PRP] ADVP
1.00	PP	→	PBAR SBAR
1.00	PBAR	→	P
1.00	P	→	before
1.00	ADVP	→	quite a lot
1.00	V[SUBCAT4,ASP]	→	had
1.00	V[SUBCAT7,PRP,COP]	→	been
1.00	V[SUBCAT2,PRP]	→	drinking
0.50	V[SUBCAT2]	→	visited
0.50	V[SUBCAT2]	→	saw
1.00	V[SUBCAT1]	→	visited
1.00	V[SUBCAT5]	→	saw
1.00	DT	→	the
0.50	ADJ	→	famous
0.50	ADJ	→	Australian
0.33	N	→	ADJ N
0.33	N	→	woman
0.33	N	→	doctor

Fig. 12. Grammar for NP/S and NP/Z ambiguities.

of the effect in the NP/S ambiguity (1a). This asymmetry has been interpreted in terms of the *Theta Reanalysis Constraint* of Pritchett (1988) and others. Sturt and colleagues suggest that the destructive nature of the reanalysis required in the NP/Z but not the NP/S ambiguity can also explain the asymmetry. Yet another alternative is that the magnitude of information transacted at the disambiguating point is different across the two sentences.

To illustrate this alternative, consider the grammar in Fig. 12. The grammar in Fig. 12 generates both of the sentences given in (1). At the disambiguating word "had" in (1a), about 3.45 bits are transacted. In (1b) at the same disambiguating word "had" about 8.79 bits are transacted. This demonstrates that the reading time asymmetry between the NP/S and NP/Z cases can be modeled as an information processing asymmetry.[3]

---

[3] As pointed out to me by William Badecker, this grammar does not literally cover the unambiguous versions of Sturt *et al.'s* (1999) stimuli; various extensions to all four classes of stimuli are possible, but space constraints preclude a full discussion of the alternatives here.

Subject and Object Relatives

The contrasting difficulty presented by subject and object relative clauses is an established processing asymmetry (see references in Gibson, 1998). From this point on, a single GPSG-type grammar, shown in Fig. 13, will be adopted.

Taking reading time to be proportional to entropy reduction, this gram-

0.20	NP	$\rightarrow$	SPECNP NBAR
0.40	NP	$\rightarrow$	I
0.40	NP	$\rightarrow$	John
1.00	SPECNP	$\rightarrow$	DT
0.50	NBAR	$\rightarrow$	NBAR S[+R]
0.50	NBAR	$\rightarrow$	N
1.00	S	$\rightarrow$	NP VP
0.87	S[+R]	$\rightarrow$	NP[+R] VP
0.13	S[+R]	$\rightarrow$	NP[+R] S/NP
1.00	S/NP	$\rightarrow$	NP VP/NP
0.50	VP/NP	$\rightarrow$	V[SUBCAT2] NP/NP
0.50	VP/NP	$\rightarrow$	V[SUBCAT3] NP/NP PP[to]
0.33	VP	$\rightarrow$	V[SUBCAT2] NP
0.33	VP	$\rightarrow$	V[SUBCAT3] NP PP[to]
0.33	VP	$\rightarrow$	V[SUBCAT4] PP[for]
0.33	V[SUBCAT2]	$\rightarrow$	met
0.33	V[SUBCAT2]	$\rightarrow$	attacked
0.33	V[SUBCAT2]	$\rightarrow$	disliked
1.00	V[SUBCAT3]	$\rightarrow$	sent
1.00	V[SUBCAT4]	$\rightarrow$	hoped
1.00	PP[to]	$\rightarrow$	PBAR[to] NP
1.00	PBAR[to]	$\rightarrow$	P[to]
1.00	P[to]	$\rightarrow$	to
1.00	PP[for]	$\rightarrow$	PBAR[for] NP
1.00	PBAR[for]	$\rightarrow$	P[for]
1.00	P[for]	$\rightarrow$	for
1.00	NP[+R]	$\rightarrow$	who
0.50	DT	$\rightarrow$	the
0.50	DT	$\rightarrow$	a
0.17	N	$\rightarrow$	editor
0.17	N	$\rightarrow$	senator
0.17	N	$\rightarrow$	reporter
0.17	N	$\rightarrow$	photographer
0.17	N	$\rightarrow$	story
0.17	N	$\rightarrow$	ADJ N
1.00	ADJ	$\rightarrow$	good
1.00	NP/NP	$\rightarrow$	ϵ

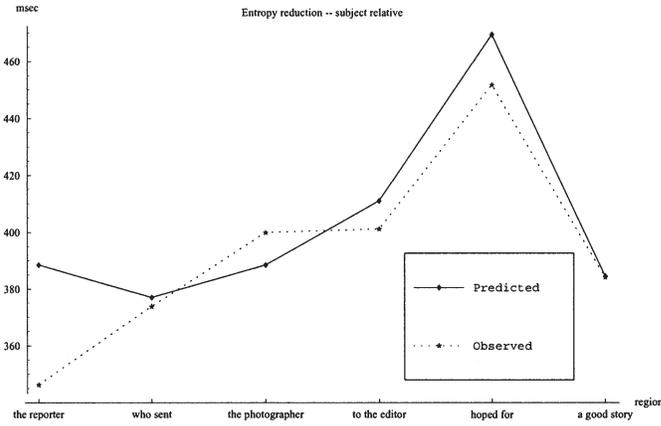**Fig. 13.** PCFG for center-embedding and Grodner *et al.* (2000) subject-object stimuli.

Fig. 14. Subject relative (observed data from Grodner *et al.* [2000]).

mar derives the subject/object asymmetry—at least in rough outline. A
reading time peak is predicted on the main verb, along with a smaller peak
on the embedded verb of just the object relative.

The graphs in Fig. 14 and 15 show the result of a regression of the pre-
dictions derived using the grammar of Fig. 13 to the average regional read-
ing times measured by Grodner, Watson, and Gibson (2000). The
correlation coefficient is 0.49, $p < .01$.

At the embedded verb in the object- (but not the subject-) relative,
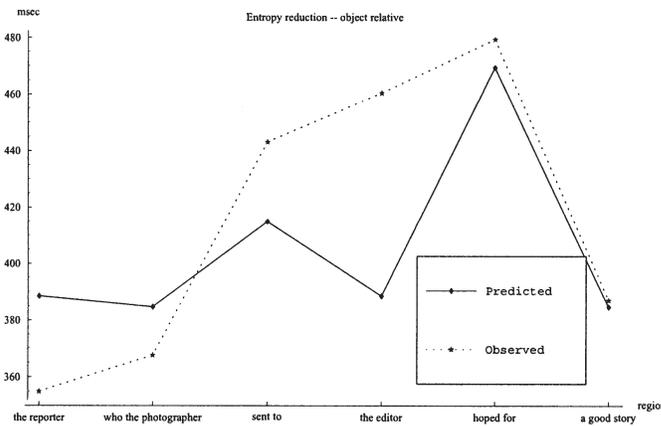entropy is reduced by about 10 bits. This happens in the object relative (see

Fig. 15. Object relative (observed data from Grodner *et al.* [2000]).

Fig. 15) because the comprehender can determine at "sent" that there will be no recursive modification of the noun phrase "the photographer." By contrast, in the subject relative (see Fig. 14), modification of the initial noun phrase "the reporter" has already been signaled by "who," a word that opens up at least as many possibilities as it resolves. There is no overt noun phrase available for modification; therefore the verb does not serve to disconfirm this possibility and entropy is not reduced in the same way. The more general claim, if the grammar in Fig. 13 is accurate, is that subject relative clauses are read more quickly on the embedded verb because the human sentence processing mechanism does not have to cope with the possibility, at that point, of recursive modification.

Center-Embedding

It has been known at least since Yngve (1960) that center-embedding induces processing difficulty, quite apart from any apparent ambiguity. Following Gibson (1998, 2000), observe that the sentences in (2) are increasingly center-embedded

(2) a. The reporter disliked the editor.
    b. The reporter [s' who the senator attacked] disliked the editor.
    c. The reporter [s' who the senator [s' who John met] attacked] disliked the editor.

Using the same GPSG-type grammar (see Fig. 13), the increasing processing difficulty associated with center-embedded sentences can be modeled by the increasing total amount of information conveyed. For the sentences in (2), these values are tabulated in Fig. 16. Whereas 47 bits are transacted processing a sentence like (2c), only 24 bits are needed for a right-branching sentence of the same length such as example (3).

(3) John met the senator who attacked the reporter who disliked the editor.

This demonstrates that the correctness of the prediction is not trivially due to the increasing length, but rather to the increasing level of embedding of

level of embedding	total entropy reduction
0	20.52
1	38.37
2	47.08

Fig. 16. Total information transacted at increasing levels of center-embedding.

the sentences involved. In fact, the final three verbs in the center-embedded sentence (2c) convey much more structural information than in their right-branching counterpart. This is because each verb resolves a structural decision between completion and continued modification of a noun phrase, a disambiguating role they do not play in the right-branching control.

CONCLUSION

The burden of the demonstrations in the last section was not to provide a tighter fit to the empirical data than was previously available. Rather, the demonstrations are meant to suggest that a range of cognitive load phenomena that have been extensively studied in sentence processing may have an intriguing explanation at a very low level: a level whose existence follows from viewing parsing as a computation in which the ambiguity of a grammar is systematically reduced as more words are processed. The proposal is that human readers are able to identify combinatory relationships in sentences because at each word they are performing the maximum amount of disambiguation, an amount of work proportional to the information conveyed by the word.

To compute the information conveyed, the procedure in Fig. 4 was used in conjunction with a symbolic, top-down parser. The operation of this mechanism is not proposed as a cognitive process; it only serves to calculate the consequences of the theoretical claim. In fact, the issue of what cognitively (or neurally) plausible processing architectures actually reduce entropy as specified here is very much open. It may be that the entropy reduction for individual parser actions can be calculated and used in a structure-building cognitive model. Such a model might operate by greedily reducing as much entropy as possible, backtracking when blocked. However, the work surveyed at the beginning of the paper suggests that an array of diverse connectionist alternatives is also possible. If Harmony Theory (Smolensky, 1986) can be applied to one or another of these alternatives, it ought to be possible to derive entropy-reducing probabilistic processing from the Harmony values of the analyses considered within a parser state. To the extent that these values reflect substantive linguistic properties, a connectionist "constraint-based" approach may offer the hope of breaking the endless circle of appealing to frequency as an explanation for performance.

APPENDIX A: CONSISTENCY OF PCFGS

It was observed early on by Grenander (1967) and Ellis (1969) that normalizing the rules of a probabilistic context-free grammar is insufficient to ensure that the defined probability model is a proper one. The probabil-

$$\begin{array}{lll} \frac{2}{3} & S & \rightarrow & S\ S \\ \frac{1}{3} & S & \rightarrow & a \end{array}$$

Fig. 17. Inconsistent PCFG.

ity model defined by a PCFG is *consistent* if the probability assigned to all sentences in the language of the grammar sums to 1.0.

Consider the normalized rules in Fig. 17. Taking S as the start symbol, and lowercase 'a' as the only terminal, this grammar generates the language a^n. However, the probability assignment to the space of strings of a only adds up to ½. Intuitively, the reason is that it's more likely that the self-duplicating S rule will be selected, rather than the derivation-ending preterminal S rule. So some derivations, the infinite ones, go on forever, sapping away probability mass from those that do end. The grammar is inconsistent.

A fully satisfactory consistency condition for PCFGs relies on a view of the derivation process as a branching process (Harris, 1963). Its statement requires the concept of an expectation matrix. The expectation matrix A of a grammar is a square matrix indexed by nonterminal-symbols where if each entry α_{ij} contains the sum of probabilities that the ith symbol is rewritten as the jth symbol. A represents the "fertility" of each grammar symbol as regards each possible kind of daughter. A grammar is consistent if each kind of child is destined to eventually die out, that is, be rewritten by only terminal symbols. This property holds if the largest eigenvalue (the "spectral radius") of the grammar's expectation matrix is less than 1.0.

REFERENCES

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Billot, S., & Lang, B. (1989). The structure of shared forests in ambiguous parsing. In *Proceedings of the 1989 Meeting of the Association for Computational Linguistics.* Vancouver.

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

Charniak, E., & Goldman, R. P. (1993). A bayesian model of plan recognition. *Artificial Intelligence, 64,* 53–79.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2,* 113–124.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge MA: MIT Press.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* John Wiley and Sons.

Crain, S., & Fodor, J. D. (1985). How can grammars help parsers? In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational and theoretical perspectives* (pp. 94–127). Cambridge: Cambridge University Press.

Den, Y., & Inoue, M. (1997). Disambiguation with verb-predictability: Evidence from

Japanese garden-path phenomena. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 179–184). Lawrence Erlbaum.

Ellis, C. A. (1969). *Probabilistic languages and automata.* Unpublished doctoral dissertation, University of Illinois, Urbana.

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

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6,* 205–254.

Frazier, L., & Clifton, C., Jr. (1996). *Construal.* Cambridge, MA: MIT Press.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37,* 58–93.

Gazdar, G., Klein, E., Pullum, G., & Sag, I. (1985). *Generalized phrase structure grammar.* Cambridge, MA: Harvard University Press.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown.* Unpublished doctoral dissertation, Pittsburgh, PA: Carnegie Mellon University.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68,* 1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (Eds.), *Image, language, brain.* Cambridge, Massachusetts: MIT Press.

Gibson, E., & Pearlmutter, N. J. (1998). Constraints on sentence processing. *Trends in Cognitive Sciences, 2,* 262–268.

Grenander, U. (1967). *Syntax-controlled probabilities* (Tech. Rep.). Providence, RI: Brown University Division of Applied Mathematics.

Grodner, D., Watson, D., & Gibson, E. (2000). Locality effects on sentence processing. In *Thirteenth Annual CUNY Conference on Human Sentence Processing.* San Diego, CA. (Talk presented at CUNY 2000).

Harris, T. (1963). *The theory of branching processes.* New York: Springer-Verlag.

Huang, T., & Fu, K. (1971). On stochastic context-free languages. *Information Sciences, 3,* 201–224.

Inoue, M., & Den, Y. (1999). *Influence of verb-predictability on ambiguity resolution in Japanese.* Poster presented at the 16th Annual Meeting of the Japanese Cognitive Science Society.

Jensen, F. V. (1996). *An introduction to Bayesian Networks.* London: University College London Press.

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

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition.* Upper Saddle River, NJ: Prentice-Hall.

Kurtzman, H. S. (1985). *Studies in syntactic ambiguity resolution.* Unpublished doctoral dissertation, MIT.

Lang, B. (1974). Deterministic techniques for efficient non-deterministic parsers. In J. Loeckx (Ed.), *Proceedings of the 2nd Colloquium on Automata, Languages and Programming* (pp. 255–269). Saarbrücken.

Lang, B. (1988). Parsing incomplete sentences. In *Proceedings of the 12th International Conference on Computational Linguistics* (pp. 365–371). Budapest.

Legendre, G., Miyata, & Smolensky, P. (1990a). Harmonic grammar—a formal multilevel connectionist theory of linguistic well-formedness: An application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 884–891). Cambridge MA: Erlbaum.

Legendre, G., Miyata, Y., & Smolensky, P. (1990b). Harmonic grammar—a formal multilevel connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. (pp. 388–395). Cambridge MA: Erlbaum.

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

Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

McClelland, J., & St. John, M. (1989). Sentence comprehension: A PDP approach. *Language and Cognitive Processes, 4,* 287–336.

McClelland, J. L., & Kawamoto, A. H. (1986). Mechanisms of sentence processing: Assigning roles to constituents of sentences. In *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 272–325). Cambridge, MA: MIT Press.

Mitchell, D. C., Cuetos, F., Corley, M. M., & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research, 24,* 469–488.

Narayanan, S., & Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society.* University of Wisconsin–Madson.

Narayanan, S., & Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading time in sentence processing. In T. G. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in neural information processing 14.* Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Meteo, CA: Morgan Kaufmann.

Pritchett, B. (1988). The grammatical basis of language processing. *Language, 64,* 539–576.

Rodriguez, P. F. (1999). *Mathematical foundations of simple recurrent networks in language processing.* Unpublished doctoral dissertation, University of California–San Diego.

Rohde, D. L. (2002). *A connectionist model of sentence comprehension and production.* Unpublished doctoral dissertation, Carnegie Mellon University.

Rumelhart, D. E., McClelland, J., & PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379–423, 623–656.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Steedman, M. (1999). Connectionist sentence processing in perspective. *Cognitive Science, 23,* 615–634.

Steedman, M. (2000). *The syntactic process.* Cambridge, MA: MIT Press.

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

Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language, 40,* 136–150.

Suppes, P. (1970). Probabilistic grammars for natural language. *Synthese, 22,* 95–116.

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

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

Tabor, W., & Tanenhaus, M. K. (2001). Dynamical systems for sentence processing. In *Connectionist psycholinguistics: Capturing the empirical data.* Westport, CT: Ablex.

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

Tanenhaus, M. K., & Trueswell, J. C. (1995). Sentence comprehension. In J. L. Miller & P. D. Eimas (Eds.), *Speech, language and communication* (Vol. 11, 2nd ed., 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., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic disambiguation. *Journal of Memory and Language, 33,* 285–318.

Yngve, V. H. (1960). A model and an hypothesis for language structure. In *Proceedings of the American Philosophical Society* (Vol. 104, pp. 444–466). Philadelphia.