Journal of Psycholinguistic Research, Vol. 29, No. 6, 2000

Wide-Coverage Probabilistic Sentence Processing

Matthew W. Crocker[1,2] and Thorsten Brants[1]

This paper describes a fully implemented, broad-coverage model of human syntactic processing. The model uses probabilistic parsing techniques, which combine phrase structure, lexical category, and limited subcategory probabilities with an incremental, left-to-right "pruning" mechanism based on cascaded Markov models. The parameters of the system are established through a uniform training algorithm, which determines maximum-likelihood estimates from a parsed corpus. The probabilistic parsing mechanism enables the system to achieve good accuracy on typical, "garden-variety" language (i.e., when tested on corpora). Furthermore, the incremental probabilistic ranking of the preferred analyses during parsing also naturally explains observed human behavior for a range of garden-path structures. We do not make strong psychological claims about the specific probabilistic mechanism discussed here, which is limited by a number of practical considerations. Rather, we argue incremental probabilistic parsing models are, in general, extremely well suited to explaining this dual nature—generally good and occasionally pathological—of human linguistic performance.

KEY WORDS: probabilistic parsing; frequency; Markov models.

INTRODUCTION

Theories of human sentence processing have largely been shaped by the study of pathologies in human sentence processing. The principles and parsing mechanisms that have been proposed are primarily directed at explaining the difficulty people have in comprehending particular structures that are ambiguous or memory intensive. While often insightful, this approach diverts attention of psycholinguists from the remarkable, yet often ignored, fact that people are, in reality, extremely accurate and effective in understanding the vast majority of utterances they encounter. That is to say, while pathologies

Both authors acknowledge the support of the DFG funded SFB (No. 378): "Resource Adaptive Cognitive Processes."

[1] Department of Computational Linguistics, Universität des Saarlandes, Saarbücken, Germany.
[2] To whom all correspondence should be addressed. email: {crocker, thorsten}@coli.uni-sb.de.

are extremely useful in exploring the boundaries of human performance and testing the predictions of particular mechanisms, this is only truly of value in the context of a concrete model of how people process language in general.

It is, therefore, not surprising that no existing model of human parsing attempts to account for both general human performance, on “garden-variety” language, and pathological behavior observed for particular ambiguities, i.e., garden-path sentences. In this paper, we argue for the importance of studying the behavior of robust, accurate, and broad-coverage parsing systems as models of human performance. The performance of the human sentence processor in dealing with the complexity, ambiguity, and noise, which pervades the linguistic environment suggests a mechanism that is extremely well adapted to its task. Computational systems that attempt to approach such coverage and accuracy require relatively powerful techniques. It is, therefore, far from clear how most extant psychological models, which are founded on assumptions of highly restricted parsing architectures, can possibly be scaled up to explain what can only be described as the exceptional standard of human performance.

We present the results of experiments conducted using the *incremental cascaded Markov model* (ICMM), a psychological model of parsing which is based on the broad coverage statistical parsing techniques developed by Brants (1999b). ICMM is consistent with accounts of human language processing that advocate probabilistic mechanisms for parsing and disambiguation (e.g. Jurafsky, 1996; MacDonald Perlmutter, & Seidenberg, 1994; Tanenhaus Spivey-Knowlton, & Hanna, 2000; Corley & Crocker, 2000). ICMM is a maximum-likelihood model, which combines stochastic context free grammar with a generalization of the hidden Markov models. The present work can be seen as a natural extension of the Statistical Lexical Category Model (Corley & Crocker, 2000), which posits a hidden Markov model-based account of human lexical category disambiguation. ICMM extends the use of Markov models from category disambiguation to full parsing, using layered, or *cascaded*, Markov models to select the most likely syntactic analyses for a given input (Brants, 1999a). To investigate psychological plausibility of the model, it has been adapted to process utterances incrementally, selecting only a subset (beam) of preferred syntactic analyses. It is important to note that restricting probabilistic parsers in this way has been separately shown to have virtually no detrimental effect on the accuracy levels for such parsers (Brants & Crocker, 2000).

As with the majority of broad-coverage, probabilistic parsers, ICMM is based on a chart-parsing algorithm, as this provides a natural way to compute all the possible structures, which are then assigned a probability, with low probability structures being pruned. It is important to clarify that we are not claiming particular plausibility for such mechanisms here, rather we are defending the general success of probabilistic models, which we assume can be associated with more psychologically justifiable models of structure building.

We begin with a brief review of probabilistic models of syntactic processing and their motivation. In particular, we observe that none of the models address the issues of general, as well as pathological, linguistic performance. We then give a description of ICMM, before presenting several simulations of the system, showing how a range of observed psycholinguistic behaviors is accounted for. In particular, we consider noun-verb category ambiguities, *that* ambiguities, and reduced relative clauses. In the final simulation, we also explain how the model accounts for the experimental findings of Pickering, Traxler, & Crocker (2000), which seemingly contradict the predictions of a pure maximum-likelihood model in NP/S complement ambiguities.

PROBABILISTIC MODELS OF SENTENCE PROCESSING

Recent research in psycholinguistics has placed increased emphasis on the role of probabilistic mechanisms (see, e.g., Seidenberg, 1997). We suggest the development of probabilistically based models of human sentence processing is motivated based on the following.

Empirical

There is strong and wide ranging psycholinguistic evidence that the human language processor is sensitive to the frequency of lexical alternatives: Duffy, Morris, and Rayner (1988) demonstrated effects of frequency on word sense disambiguation. Corley and Crocker (2000) demonstrate how a statistical model of category disambiguation, when trained on a corpus, successfully models a number of observed experimental findings (see also Crocker & Corley, in press, for further experimental support). Trueswell (1996) demonstrates the sensitivity of the human parser to the preferred tense for a given verb. Jurafsky (1996) motivates a probabilistic model of lexical and syntactic processes. Probabilistic models are further supported by recent corpus studies (Lapata, Keller, Schulte im Walde submitted) which suggest that corpus frequencies correlate well with subcategorization preferences observed in completion studies by Trueswell, Tanenhaus, and Kello (1993), Garnsey, Pearlmuter, Myers, and Lotockey (1997), Pickering *et al.* (2000) and others.

Computational

The use of statistical language models in computational linguistics has proved to be extremely successful in developing broad-coverage models, which can accurately estimate the most likely parse (Collins, 1996;

Ratnaparkhi, 1997). In the context of psychological modeling, Brants and Crocker (2000) have also demonstrated that the performance of probabilistic parsing models does not deteriorate, even when incremental processing and strict memory limitations are imposed.

Rational

The success of probabilistic models helps explain the *rational* nature of the human language processor, i.e., that the human parser is generally able to accurately, rapidly, and robustly recover the appropriate interpretation for the utterances it encounters. Within the framework of *Rational Analysis* (Anderson, 1991), Chater, Crocker, and Pickering (1998) motivate the use of a probabilistic framework in deriving a model of human parsing and reanalysis based on the hypothesis that the human language processor is well adapted to the problem of resolving linguistic ambiguity. Crocker and Corley (in press) also point out that probabilistic mechanisms provide highly accurate heuristic mechanisms, which are particularly well suited to modular architectures where full knowledge is not immediately available, and must be approximated.

Research in experimental and computational psycholinguistics has focussed primarily on explaining the role of probabilistic mechanisms for several well-known garden-path constructions. Constraint-based models, for example, have long argued for the importance of lexical biases in ambiguity resolution (e.g. MacDonald *et al.*, 1994; Trueswell, 1996; McRae *et al.*, 1998). The model outlined by MacDonald and colleagues is probabilistic in the sense that alternative feature values of ambiguous lexical items are associated with probabilistically determined activations (e.g., *examined* might have a higher activation as transitive, rather than intransitive). Lexical items are combined to build syntactic analyses, with the activation of each analysis being determined by the combined activation of the relevant linguistic constraints. To our knowledge, however, the model is not implemented, nor is it very transparent how probabilistic feature activations are to be acquired and combined. As a result, the model is not sufficiently well specified to make concrete predictions.

In contrast, McRae *et al.* attempt to concretely demonstrate the predictions of a model, which simultaneously combines several probabilistic constraints to resolve syntactic ambiguity using the *competition-integration* model. Crucially, however, McRae *et al.* only model the interaction of constraints in *selecting* among interpretations and do not model the parse/interpretation-building process itself. The model is interesting, however, in that the constraint activations are established empirically (using a mixture of corpus and norming studies) and constraint weights are then determined by fitting off-line completion data. The resulting model is then shown to provide a good fit of human reading time data for the same items.

While both of these models can be viewed as incorporating probabilistic constraints, there are some problems with regarding this as a truly probabilistic approach. The McRae model conflates constraints that are established using corpora with those derived (linearly) from ratings. In addition, the competition-integration mechanism only uses these "probabilities" to determine initial activation of analyses—subsequent cycling of the model changes activations in such a way that they no longer have any probabilistic interpretation. The MacDonald *et al.* model is also subject to the latter criticism.

More importantly, from the perspective of the current paper, it is unclear how such constraint-based models, will scale into a full model of sentence processing.[3] Furthermore, the competition mechanism predicts that local ambiguities in which competing analyses have similar activations (nee probabilities) will take longer to resolve. While this has been demonstrated to provide an interesting fit of human reading times for reduced relative clauses (McRae *et al.,* 1998) and several other constructions (Tanenhaus *et al.,* 2000), it is unclear whether this prediction is sustained for language processing, in general. A true probabilistic model, in contrast, makes no such prediction: the probability of analyses simply determines the ranking of interpretations at each point during processing of the utterance.

Jurafsky (1996) presents a computational model of lexical access and syntactic disambiguation, which is truly probabilistic. The model associates probabilities with various linguistic representations, including phrase structure rules and lexical valence (i.e., subcategorization). When utterances are processed, the probability of alternative structures is computed by combining the probabilities of the contributing rules and lexical entries, which are utilized in each analysis. Alternative analyses are then ranked according to their probability and those structures below a given threshold are eliminated, thus enforcing memory constraints. From a theoretical perspective, the model Jurafsky proposes is very much in the spirit of the approach we develop in this paper and demonstrates the success of probabilistic mechanisms in providing principled, unified, and predictive accounts for a range of psycholinguistic phenomena.

As with other psycholinguistic models, however, the coverage and scalability of Jurafsky's model remains unclear and certainly unproved. Indeed, to our knowledge, the only broad-coverage model of sentence processing is that of Corley and Crocker (2000). They present a model of human lexical category disambiguation that is based on a probabilistic hidden Markov model. Such models have been shown, in the general case, to be extremely accurate (Brants, 2000), while Corley and Crocker also demonstrate that such a model

[3] Indeed, this criticism can be leveled at most models of human sentence processing.

can explain a range of results concerning human processing of category-ambiguous words. This present work builds directly on their approach, but extends it beyond category disambiguation to full syntactic parsing.

CASCADED MARKOV MODELS

The basic idea of cascaded Markov models is to construct the parse tree layer by layer, first structures of depth one, then structures of depth two, and so forth. For each layer, a Markov model determines the best set of phrases. These phrases are used as input for the next layer, which adds one more layer. Phrase hypotheses at each layer are generated according to stochastic context-free grammar rules (the outputs of the Markov model) and subsequently filtered from left to right by Markov models.

Figure 1 gives an overview of the parsing model by showing the processing steps for a simple example sentence taken from the Wall Street Journal corpus (Marcus, Santorini and Marcinkiewicz, 1993). A cascaded Markov model consists of a stochastic context-free grammar and a separate Markov model for each layer (up to some maximum number of layers). The first layer resolves lexical category ambiguities by tagging each word with its most likely part-of-speech. New phrases are created at higher layers and filtered by Markov models operating from left to right. Only those hypotheses reaching a particular probability value are passed up to the next higher layer; the others are pruned.

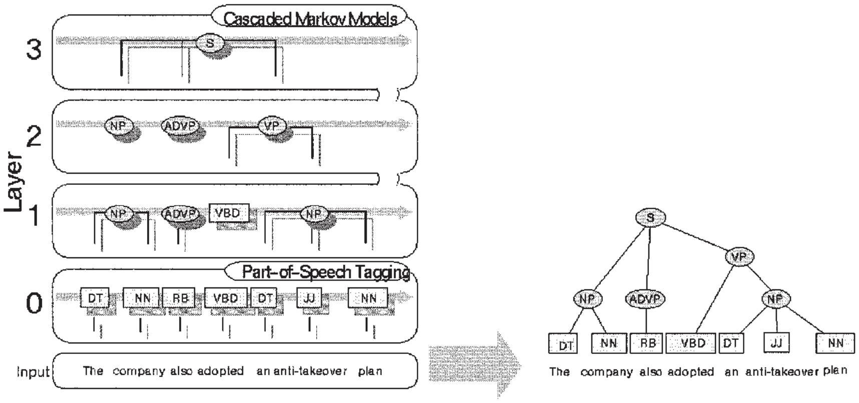


Fig. 1. The layered processing model. Starting with part-of-speech tagging (layer 0), possibly ambiguous output together with probabilities is passed to higher layers (only the best hypotheses are shown for clarity). At each layer, new phrases are added and filtered with a Markov model.

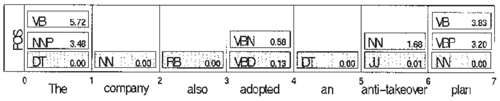


Fig. 2. The part-of-speech layer. For each word, the possible tags and their γ probabilities (negative logarithm: thus smaller values correspond to higher probabilities) are shown. For statistical part-of-speech tagging, this represents a lattice and the task is to find the optimal path from nodes 0 to 7.

The Part-of-Speech Layer

For part-of-speech disambiguation, we use the hidden Markov model approach as implemented by Brants (2000). This layer is largely similar to the psychological model proposed by Corley and Crocker (2000). This approach first retrieves, for each word, the allowed tags and their lexical probabilities from a lexicon.[4] It then selects the best sequence of tags by taking additionally contextual probabilities into account. Figure 2 shows all allowed tags for the example sentence and the negative logarithm of their γ probabilities. These result from the combination of lexical probabilities $P(word \mid tag)$ and contextual probabilities $P(tag_3 \mid tag_1 tag_2)$ [a second-order Markov model, while Corley and Crocker use a first order model: $P(tag_2 \mid tag_1)$]. Calculation of γ (or forward–backward) probabilities is described in (Rabiner, 1989). The sequence of part-of-speech tags with the highest probability is shaded gray in Figure 2.

Passing Hypotheses to the Next Layer

After having processed a layer, the best hypotheses and alternatives with high probabilities are passed to the next layer. Those alternative tags are shaded light gray in Figure 2. We employ a beam of 100, i.e., a tag is passed if its probability is at least 100th of the best tag's probability. This factor of 100 is equivalent to a difference of 2 in the negative logarithms. All tags having a value, which is, at most, 2 larger than the best one, are passed and therefore shaded light gray. All tags with a white background are ruled out at the part-of-speech layer.

Passing more than one hypothesis is advantageous in case a lower-layer model introduces an error. We increase the chance that the correct tag is among those that are passed. The higher-level model identifies the alternatives and their probabilities and can choose among them. We decide against passing only one hypothesis to the next layer because this would make it impossible for higher layers to correct errors introduced at lower

[4] If a word is not found in the lexicon, the tagger generates a probability distribution over all tags according to a statistical suffix analysis.

layers. We also decide against passing all hypotheses, because we want to keep parallelism in the model as low as possible. The empirically determined value of 100 results in an average of 1.3 tags per word passed to the first structural layer.

Generating Phrases According to a Context-Free Grammar

After having selected part-of-speech tags with high probabilities, the model consults a stochastic context-free grammar and adds new phrases to the hypothesis space. The phrase hypotheses at layer 1, for the example sentence, are shown in Figure 3. Those elements that are passed from the lower layer have a bold frame, all others are added according to the grammar. Very typical for a stochastic context-free grammar, the number of hypotheses can become quite large. This part is identical to filling the chart in context-free parsing. We just restrict the generation of new phrases to one layer.

Tagging Lattices

The hypotheses for layer 1 form a lattice, with the word boundaries being states and the phrases being edges. Selecting the best hypotheses means to find the best path from node 0 to the last node (node 7, in the example). The best path can be efficiently found with the Viterbi (1967) algorithm, which runs in time linear to the length of the word sequence. Having this view of finding the best hypothesis, processing of a layer is similar to word-lattice processing in speech recognition (cf. Samuelsson, 1997).

Two types of probabilities are important when searching for the best path in a lattice. First, these are probabilities of the hypotheses (phrases) generating the underlying terminal nodes (words). They are calculated according to a stochastic context-free grammar. The second type are context probabilities, i.e., the probability that some type of phrase follows or precedes another. The two types of probabilities coincide with lexical and contextual probabilities of a Markov model, respectively. According to a trigram model (generated from a corpus), the path in Figure 3 that is shaded dark grey is the best path in the lattice. Its probability is calculated as follows:[5]

$$\begin{aligned}P_{\text{best}} = {} & P(\text{NP} \mid \text{start}) \cdot P(\text{NP} \Rightarrow {}^* \text{ The company also}) \\ & \cdot P(\text{VBD} \mid \text{NP, ADVP}) \cdot P(\text{VBD} \Rightarrow {}^* \text{ adopted}) \\ & \cdot P(\text{NP} \mid \text{ADVP, VBD}) \cdot P(\text{NP} \Rightarrow {}^* \text{ an anti-takeover plan}) \\ & \cdot P(\text{end} \mid \text{VBD, NP})\end{aligned}$$

[5] Note that this layer incorrectly prefers to attach the adverb to the NP. However, the correct analysis is among those with high probabilities, and will be preferred at the higher layer.

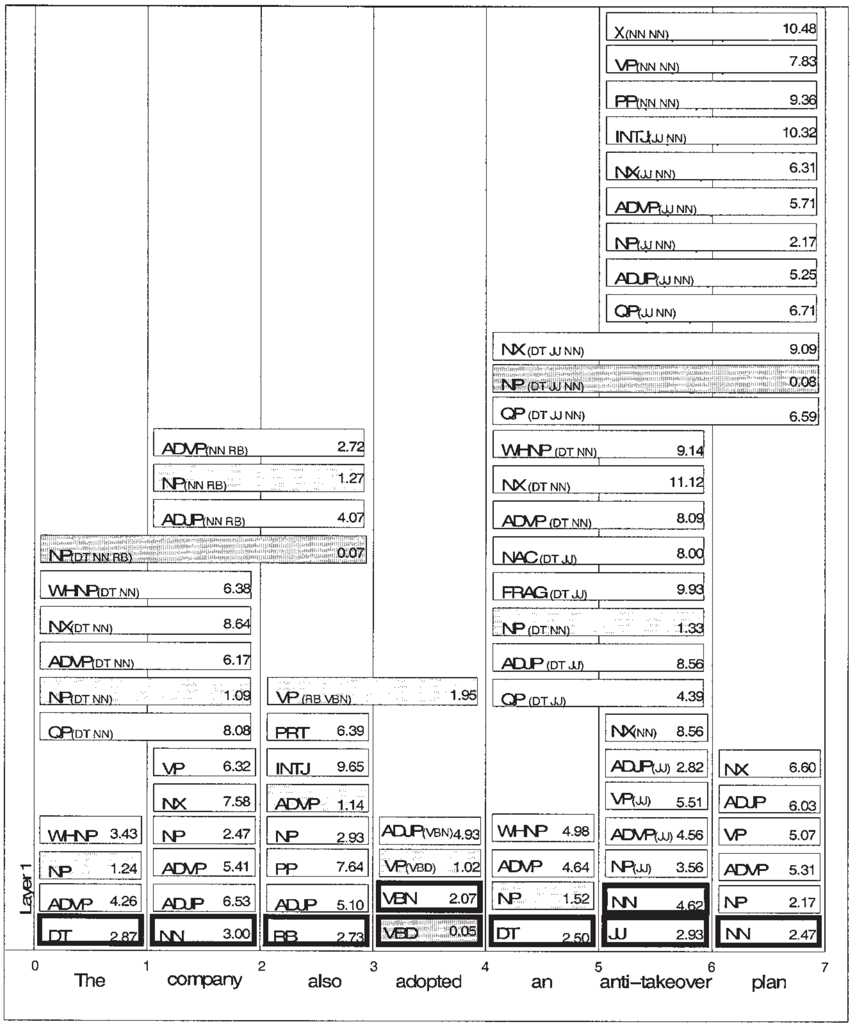


Fig. 3. Phrase hypotheses and their probabilities (negative logarithm) at layer 1. As for the part-of-speech layer, the task of the Markov model is to find the optimal path from nodes 0 to 7. Elements with a bold frame were passed from layer 0. The gray elements (11 of 68) have high probabilities and are passed to layer 2; the best path is dark gray.

The best path correctly predicts the two NPs and the ADVP. For each phrase, the γ probability (negative logarithm) is given in Figure 3. All hypotheses that are within the pre-defined beam of factor 100 are collected and passed to layer 2. In this example, we find an average of 2.7 passed hypotheses in parallel (opposed to 14.6 before filtering).[6] The presented Markov models act as *filters.* The probability of a connected structure is determined only based on a stochastic context-free grammar. The joint probabilities of unconnected partial structures are determined by using Markov models, in addition. While building the structure bottom up, parses that are unlikely, according to the Markov models, are pruned.

A modified Viterbi algorithm is used to process Markov models operating on lattices. In part-of-speech tagging, each hypothesis (a tag) spans exactly one word. Now, a hypothesis can span an arbitrary number of words and the same span can be covered by an arbitrary number of alternative word or phrase hypotheses. Using terms of a Markov model, a state is allowed to *emit a context-free partial parse tree,* starting with the represented nonterminal symbol, yielding part of the sequence of words. This is in contrast to standard Markov Models. There, states emit atomic symbols. Note that an edge in the lattice is represented by a state in the corresponding Markov model.

Figure 4 shows the part of the Markov model that represents the best path in the lattice of Figure 3. Details of calculating the best path and γ probabilities for each element are described in Brants (1999b; 2000).

Generating, Filtering, Passing

In the example, layer 1 contains 68 hypotheses and passes those 11 elements with high probabilities (shaded gray in Fig. 3) to layer 2. There again, new phrases are generated according to the stochastic grammar, filtered with a Markov model, passed to layer 3, etc. The process iterates either until a single highly ranked phrase spans the entire input or until some predefined topmost layer is reached. In the latter case, the best path represents the resulting partial parse. Proceeding with the example sentence, layer 2 would generate 161 phrase hypotheses, of which 15 are passed to layer 3. There, 70 new phrases are generated, of which 10 are passed to layer 4. Since one of them (an *S* node) spans the entire input, and has high probability, the process stops and emits the structure, as shown in Figure 1.

Incremental Cascaded Markov Models

For our investigations, cascaded Markov models are set up to run incrementally. After reading each word, hypotheses are generated at the different

[6] These are averages per word. There are 10 hypotheses on top of *The,* of which 3 are passed, 15 on top of *company* of which 3 are passed, etc.

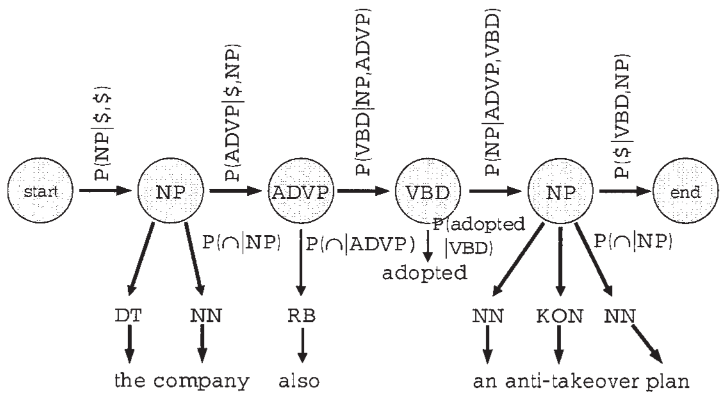


Fig. 4. Part of the Markov models for layer 1 that is used to process the sentence of Figure 3. Contrary to part-of-speech tagging, outputs of states may consist of structures with probabilities according to a stochastic context-free grammar.

layers and subsequently filtered. The original algorithm processed and finished each layer before proceeding to a higher layer. Incremental processing and filtering is a harder task since no right context is inspected. Instead, the process must hypothesize about future input.

For the incremental variant, we need to make two additional decisions: whether to filter active edges, in addition to inactive edges, and whether we should build hypotheses on inactive edges or not.

A chart-parsing process generates two types of chart entries: inactive edges, which represent complete hypothesised constituents, and active edges, which represent prefixes of hypothetical constituents. We concentrated on filtering inactive edges (recognized constituents) in the nonincremental version of our model. This was appropriate since we knew the entire input and could immediately generate all inactive edges. Now, in the incremental version, it may be advantageous to filter out some of the active edges before proceeding to the next word. This reduces memory and processing load since some of the prefixes are dynamically eliminated and need no further inspection. For our investigation, we decided to filter both active and inactive edges.

Active edges represent hypothetical constituents, which may be completed by future input. Should a higher layer already start to build new hypotheses on top of this incomplete constituent or should it wait until the lower layer constituent is completed? We chose the former, immediately starting the higher layer process. This makes processing faster since our model inherently views the different layers as parallel processes.

Parameter Estimation

A big advantage of cascaded Markov models is that they are entirely trained on corpus data. This ensures wide coverage and robustness. Transitional parameters for cascaded Markov models are estimated separately for each layer. Output parameters are the same for all layers, they are taken from the stochastic context-free grammar that is read off the treebank.

Training on annotated data is straightforward. First, we number the layers, starting with 0 for the part-of-speech layer. Subsequently, reformation for the different layers is collected.

Each sentence in the corpus represents one training sequence for each layer. This sequence consists of the tags or phrases at that layer. If a span is not covered by a phrase at a particular layer, we take the elements of the highest layer below the actual layer. Figure 5 shows the training sequences for layers 0–3, generated from the structure in Figure 1. Each sentence gives rise to one training sequence for each layer. Contextual parameter estimation is done in analogy to models for part-of-speech tagging and the same smoothing techniques can be applied. We use a linear interpolation of uni, bi-, and trigram models.

A stochastic context-free grammar is read directly off the corpus. The rules derived from the annotated sentence in Figure 1 are also shown in Figure 5. The grammar is used to estimate output parameters for all Markov models, i.e., they are the same for all layers. We could estimate probabilities for rules separately for each layer, but this would worsen the sparse data problem.

Layer	Sequence			
3		S		
2	NP	ADVP	VP	
1	NP	ADVP	VBD	NP
0	DT NN	RB	VBD	DT JJ NN

Context-free rules and their frequencies

S	→ NP ADVP VP	(1)	NP → DT JJ NN	(1)
VP	→ VBD NP	(1)	DT → *The*	(1)
NP	→ DT NN	(1)	...	
ADVP	→ RB	(1)	NN → *plan*	(1)

Fig. 5. Training material generated from the sentence in Figure 1 (right). The sequences for layers 0–3 are used to estimate transition probabilities for the corresponding Markov models. The context-free rules are used to estimate the SCFG, which determines the output probabilities of the Markov models.

MODELLING HUMAN PARSING AND REANALYSIS

Cascaded Markov models are part of a growing family of probabilistic parsing techniques developed primarily for the task of accurately and robustly find the most likely parse for naturally occurring, garden-variety, language (often defined more concretely with respect to exemplary corpora). While such probabilistic parsers, including the ICMM, are far from perfect, we suggest they provide the best available approach for robustly and accurately dealing with linguistic complexity, ambiguity, and noise (such as mild ungrammaticalities, slips of the tongue, etc.). As such, we claim that models like ICMM provide a plausible, if crude, first approximation of general human linguistic performance.

In this section we demonstrate that, in addition to obtaining good overall performance, the ICMM also successfully explains human behavior in several well-studied locally ambiguous constructions. As our claims concerning the psychological reality of the ICMM are focused on it's probabilistic disambiguation mechanism,[7] we focus here on modeling experimental results, which have explicitly manipulated likelihood. It is important to note that the following simulations are generated by the ICMM as trained on the Wall Street Journal portion of the Penn Treebank (Marcus *et al.*, 1993), and that the model has not been "tuned" in any way for these examples.[8]

Lexical Category Ambiguity

As Crocker and Corley (in press) point out, lexical category ambiguity is a significant, and frequent, problem for human language processing. Their study of the Brown corpus revealed that 10.9% of word *types* and 65.8% of word *tokens,* are category ambiguous in English. For example, words that are ambiguous between noun and verb readings are very common in English. Frazier and Rayner (1987) and MacDonald (1993) both exploited this observation in experiments which investigated noun–verb ambiguities in sentences of the following sort:

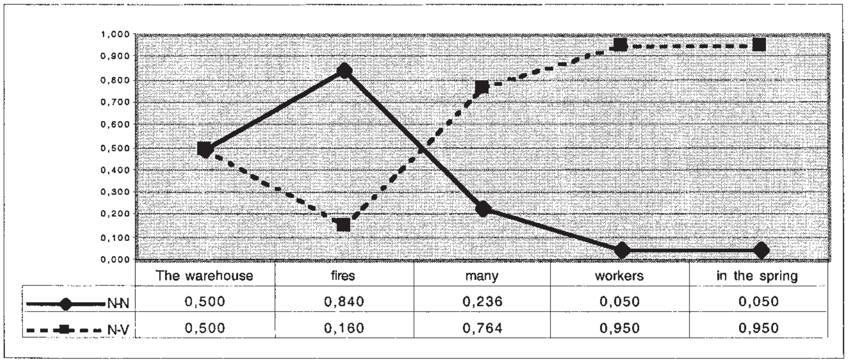
(1a) The warehouse *fires*$_{\mathrm{V}}$ many workers in the Spring.
(1b) The warehouse *fires*$_{\mathrm{N}}$ are difficult to control.

[7] That is to say, we do not make particular psychological claims concerning the underlying incremental chart parsing algorithm, for example. The only crucial property of the parser, w.r.t the probabilistic mechanism, is that it incrementally constructs all analyses at each point in processing (where most will be immediately pruned).

[8] It was necessary to use the Wall Street Journal section, instead of the more balanced Brown corpus, since only the former made available the necessary subcategory information.

Results of these studies were taken as support for a delay strategy and an interactive constraint-based view, respectively. However, neither study controlled for the frequency bias of the ambiguous word. In contrast, probabilistic models of category disambiguation (Corley & Crocker, 2000), the parsing models of Jurafsky (1996), and the model developed here, predict that lexical frequency information will be fundamental in resolving such ambiguities. Experimental findings of Crocker and Corley (in press) demonstrate that, as predicted, the category frequency bias of the ambiguous word is a fundamental determinant of how local ambiguity is initially resolved. In particular, they find that reading times in the disambiguating region immediately following an ambiguous, but noun-biased, item, like *fires*, are significantly higher when the continuation forces a verb interpretation than when it is consistent with the noun interpretation. A corresponding effect is observed when verb-biased items are noun disambiguated. Their findings indicate that, all other things being equal, the human sentence processor will initially prefer analyses, which associate an ambiguous word with its most frequently observed category.

Given that the present model incorporates a nearly identical mechanism for lexical category disambiguation to the hidden Markov model of Corley and Crocker (2000), it should not be surprising that the ICMM similarly accounts for the experimental findings. For reasons of space, we therefore only exemplify, in Figure 6, the behavior of the parser for a sentence containing noun-biased word, namely *fires,* which is subsequently disambiguated as a verb. As shown in the graph,[9] the ICCM predicts an increased reading time due to reanalysis when the disambiguating region (beginning with *many* . . .) is processed. The parser exhibits a corresponding pattern of behav-



	The warehouse	fires	many	workers	in the spring
◆ N-N	0,500	0,840	0,236	0,050	0,050
■ N-V	0,500	0,160	0,764	0,950	0,950

Fig. 6. Parse probabilities for a noun-biased item, where the continuation forces verbal reading.

[9] The probabilities shown in the graphs have been re-normalized to sum to one, so that the relative probability of the two analyses can be seen more clearly.

ior when verb-biased items are subsequently disambiguated as nouns. The behavior of the system is, therefore, consistent with the findings of Crocker and Corley.

The ICMM similarly models the effect of immediately preceding context in biasing the most likely category, as demonstrated in the experiments of Juliano and Tanenhaus (1993). In particular, they show that the preferred category assignment for the ambiguous word *that*, is as a determiner, when it occurs in the sentence initially and as a complementizer when it appears postverbally, as illustrated in the following sentences:

(2a) The lawyer insisted *that*$_{\text{Comp}}$ experienced diplomats would be very helpful.
(2b) *That*$_{\text{Det}}$ experienced diplomat would be very helplul to the lawyer.

For reasons of space, we do not elaborate here on precisely how the ICMM simulates the findings of Juliano and Tanenhaus (1993). Rather, the reader is referred to Corley and Crocker (2000) for a detailed explanation, which also holds for the system described here.

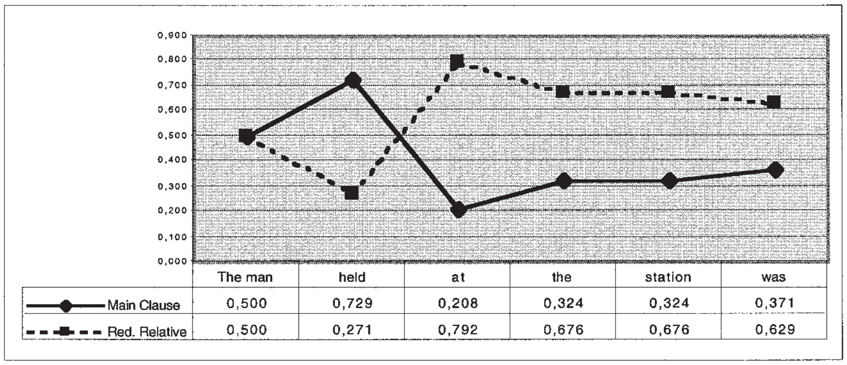
Reduced Relatives

Garden-path effects in reduced relative clauses have long been taken as strong support for the importance of purely syntactic disambiguation strategies (see e.g., Ferreira & Clifton, 1986, and references cited therein). A number of recent studies, however, have convincingly demonstrated the important role of other linguistic knowledge, such as lexical, lexico-syntactic, thematic, and discourse factors, in resolving such ambiguities (see e.g., Merlo & Stevenson, 2000; Altmann & Steedman, 1988; McRae *et al.* 1988; Tanenhaus *et al.* 2000). MacDonald (1994), for example, demonstrated that the transitivity preference of the ambiguous verb, combined with the cue provided by a following prepositional phrase following the verb, conspire to facilitate the necessary reanalysis to the reduced relative clause interpretation.

Because of the sparseness of data for the precise materials used by MacDonald, we use slightly different items in the present simulation of MacDonald's findings. In particular, we consider the sentences shown in (3), where (3a) corresponds with MacDonald's transitively biased items, while (3b) is used to represent the instransitively biased materials.

(3a) The man *held*$_{\text{Trans}}$ at the station was arrested.
(3b) The man *raced*$_{\text{Intrans}}$ to the station was arrested.

The simulation shown in Figure 7, illustrates how, for the transitive items like (3a), the parser is able to immediately switch to the correct reduced relative analysis as soon as the prepositon following the ambiguous verb is processed. This results from the low probability given to the alternative, main



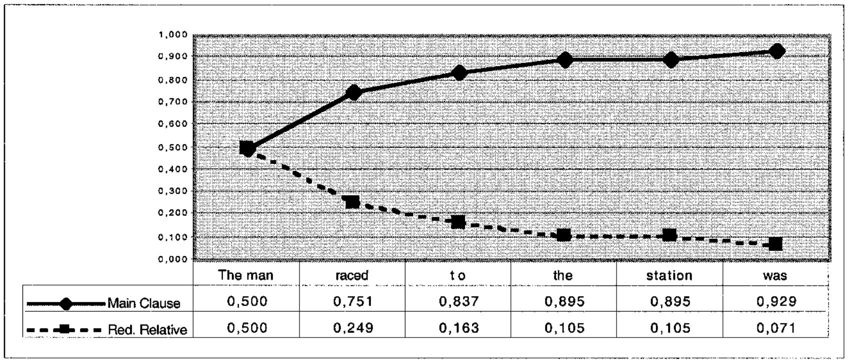
	The man	held	at	the	station	was
Main Clause	0,500	0,729	0,208	0,324	0,324	0,371
Red. Relative	0,500	0,271	0,792	0,676	0,676	0,629

Fig. 7. Parse probabilities or the reduced relative ambiguity for a transitive-biased verb like *held.*

clause reading, since the verb would need to be interpreted with its lower probability intransitive frame. Figure 8, in contrast, shows that for intransitive items like (3b), the prepositional phrase provides no such cue. The intransitive VP of the main clause analysis is consistent with the verbs preferred usage.

In related work, McRae *et al.* (1998) argue for a fully constraint-based model of sentence processing, in which all relevant linguistic constraints are immediately recruited to resolve ambiguity. Specifically, he uses the competition-integration model (Spivey-Knowlton, 1996) to fit off-line biases for several linguistic constraints to reading times for reduced relative-clause sentences, such as those in example (4).

(4a) The crook arrested by the detective was guilty of taking bribes.
(4b) The cop arrested by the detective was guilty of taking bribes.



	The man	raced	to	the	station	was
Main Clause	0,500	0,751	0,837	0,895	0,895	0,929
Red. Relative	0,500	0,249	0,163	0,105	0,105	0,071

Fig. 8. Parse probabilities of the reduced relative ambiguity for an intransitive-biased verb like *raced.*

For present discussion let us consider only those four constraints, which are postulated to come into play when the ambiguous verb is encountered:

1. *Main clause bias:* the overall bias to build a MC over RR due to its higher frequency
2. *Verb-tense/voice bias:* the lexical frequency bias of the verb to be used in either the simple past or past-participle form
3. *by-bias:* the support for building a RR which result from the parafoveally observed *by*-phrase
4. *Thematic fit:* the support for MC contributed by good agents versus support for RR contributed by poor agents

McRae *et al.* argue that a constraint-based model (as approximated using the competition-integration model with all relevant constraints immediately available) provides a better fit of on-line processing than a modular, garden-path model (which is implemented by delaying all but the first constraint in the list above). It is interesting to note, however, that the present model can also be viewed as modular, in that no postsyntactic constraints are made available during the initial stages of parsing. In contrast with the garden-path model, however, the ICMM does make use of both lexical and syntactic probabilities. Indeed this observation highlights the fact that probabilistic mechanisms are equally consistent with both modular and interactive architectures. The ICMM, therefore, effectively includes both the first and second constraint above, as well as the transitivity bias of the verb (which McRae *et al.* omit). Furthermore, while the preposition is not modeled parafoveally (the third constraint above), the simulation in Figure 7 demonstrates clearly how the information supplied by the preposition is used immediately to revise the probabilities of the alternatives. We would, therefore, expect probalistic, but nonetheless modular, models like the ICMM to fit the on-line reading data of McRae *et al.* better than their “garden-path” model. It is also important to note that while McRae *et al.* set the “off-line” parameters individually, the ICMM learns all parameters via a uniform, automatic, and mathematically well-founded training procedure. Furthermore, there is no separate “fitting” of weights for the individual constraints. As a result, such truly probabilistic models make stronger and clearer predictions and, more importantly, do so in a model of processing that actually explains how probabilistic mechanisms are used in *building* and ranking alternative interpretations.

NP-S Complement Ambiguity

In the final simulation, we consider evidence that has recently been used to argue against likelihood-based approaches. The NP/S complement ambiguity arises when a verb's subcategorization requirements can be fulfilled by both NP or bare S complements. As illustrated in example (5), at the point of processing an NP, immediately following an ambiguous NP/S-complement verb, comprehenders must decide whether to interpret the NP as a direct object or embedded subject.

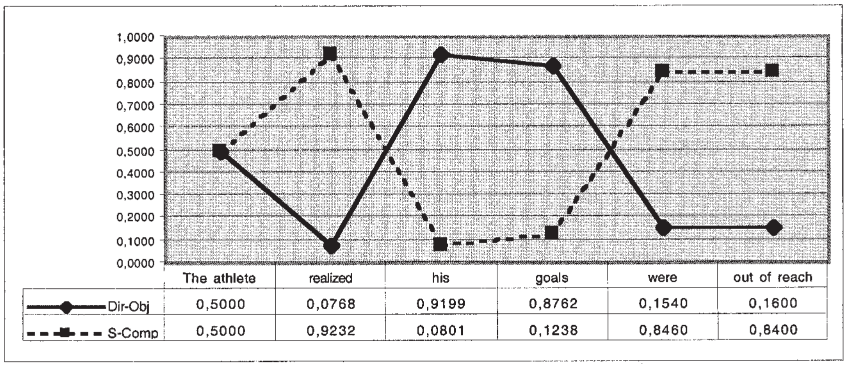
(5a) The athlete realized $[_{NP}$ his goals] at the Olympics
(5b) The athlete realized $[_{S}[_{NP}$ his goals] were out of reach]

Probabilistic ambiguity resolution mechanisms naturally predict that a primary determinant of the preferred structure will be the subcategorization bias of the verb (see e.g., Garnsey *et al.*, 1997). Recent experiments by Pickering *et al.*, 2000), however, provide convincing evidence that people initially attempt the direct object attachment for such ambiguities, even for S-biased verbs. As they point out, their result stands in direct opposition to the predictions of a strict likelihood model (i.e. models in which likelihood estimates correspond to the most preferred structures).

While the present model is likelihood based, the calculation of probabilities for a particular (partial) analysis, is *not* based upon the frequency with which that *analysis* has been seen before. Rather, the probability of an analysis is (imperfectly) *approximated* by computing the product of the probabilities of the individual rules used in the analysis, as made clear earlier. As a result, the S-complement, analysis, which requires an additional phrase structure rule to complete the attachment, will tend to have a lower probability than the direct-object analysis. This occurs despite the fact that the verb's subcategorization bias of the verb will favor the appropriate VP rule (i.e.,VP → S, in this case). While this method of calculating probabilities might be criticized for not assigning sufficiently accurate likelihoods to particular structures, it can be thought of as implementing a preference for "simpler" structures.

Figure 9 shows the probabilities assigned by the parser to the competing analyses during processing. As we can see, the verb is initially attached with its more likely S-complement subcategorization frame. However, as soon as it is followed by the (left frontier of) a noun phrase, it assigns a higher probability to the competing (and simpler) direct-object analysis. This is sustained until the disambiguating region, when the S-complement analysis is then reassigned a higher probability. The ICMM, therefore, predicts a preference for initially attaching the NP as a direct object, despite the S-complement bias of the verb.

The parser's behavior is thus largely consistent with the findings of Pickering *et al.* which demonstrated an increased reading time effect on the postverbal NP, when it was an implausible direct object (suggesting readers initially attempt and interpret it as a direct object and must immediately rean-



	The athlete	realized	his	goals	were	out of reach
◆ Dir-Obj	0,5000	0,0768	0,9199	0,8762	0,1540	0,1600
■ S-Comp	0,5000	0,9232	0,0801	0,1238	0,8460	0,8400

Fig. 9. Parse probabilities of the NP/S ambiguity for an S-biased verb.

alyze). In conditions where the direct object reading was plausible, they found and increased reading time in the disambiguating region, which the parser predicts as a result of switching the from the previously favored direct-object analysis to the now unambiguous S-complement analysis. It could be argued that the ICMM also acts a reanalysis effect at the beginning of the ambiguous NP (when the preference switches from S to NP complement). However, we would expect any such effect to very small, since it only entails reranking of the verbs subcategorization preference, and not any structural reanalysis. Pickering *et al.* found no evidence of such an effect.

DISCUSSION

This paper has presented a probabilistic model of parsing that is designed to achieve good performance on general language processing, while also explaining a number of pathological behaviors in processing local ambiguities. Our claims regarding the psycholinguistic plausibility of the presented models are primarily restricted to the probabilistic disambiguation mechanism, in which alternative analyses are ranked by the parser according to their estimated likelihood, with low probability analyses being discarded. For full discussion of general performance, the reader is referred to related work by Brants and Crocker (2000). Summarized briefly, Brants and Crocker present detailed results showing that the enforcement of strict incremental processing, combined with substantial pruning of low probability structures, has virtually no adverse effect on the accuracy of an SCFG-based parser, similar to the one presented here. In addition to being able to reduce the memory requirements to 1% of the total search space, the enforcement of memory restrictions also leads to a reduction in the average parse time by up to two orders of magnitude. In addition to showing the sustained accuracy of incremental, resource-bound probabilistic parsers, their result is important in countering the possible criticism that probabilistic parsers are too powerful and resource intensive to be considered as the basis of a cognitively plausible model.

Constraints imposed by our desire to build a broad-overage model of sentence processing (i.e., one that can be trained on, and tested against, available parsed corpora of naturally occurring language), entail a probabilistic model, which is easily considered naive in several respects. The lexicon contains only words and their possible syntactic category (and associated probabilities). The grammar, which is determined directly from the trees in the parsed treebank corpus, also reflects the aims of practical linguistic coverage over fidelity to any sophisticated linguistic theory. The present work should therefore be seen as complementary to the work of Jurafsky (1996): where Jurafsky gives up broad-coverage implementability in favor of a richer, more psychologically likely account, we trade-off in the opposite direction. However, we suggest that even our less sophisticated probabilistic model provides a compelling explanation for a range of observed human processing phenomena.

As we point out, there are number of interesting points that emerge in comparing our probabilistic model of syntactic processing, with constraint-based models that also exploit probabilistic constraints. We suggest that our approach is methodologically superior on several grounds. ICMM relates the probabilistic mechanism directly to the representation building processes of the parser and always manipulates true probabilities, rather than converting them to activations that subsequently lose any transparent probabilistic interpretation. Furthermore, we have a clearly defined and uniform training procedure that determines all the parameters of the model similarly. This means the combination of these probilities in determining the probability of a particular analysis also has a clear and well-defined probabilistic interpretation. Equally, no separate fitting of "constraint weights" is needed, thereby eliminating the possibility of fitting the model to process only a single construction well. In ICMM, parameters are estimated from large corpora, as an approximation of human linguistic experience, and the same parameter values are used in processing all utterance types. One area in which the competition-integration model is superior, is that it makes relatively clear (and, therefore, potentially falsifiable) predictions about actual observed reading times, while probabilistic models only give a ranking. A mapping function from probabilistic parser behavior to reading times remains an interesting and open area of inquiry.

We should also be clear that there is still much scope for research into the precise nature of the probabilistic human sentence processor. Our simulation of the NP/S complement ambiguity perhaps best exemplifies this. We noted that the ICMM accounts for observed behavior because of its bias toward simpler structures in estimating probabilities. Models that condition probabilities on richer lexical and structural contexts might no longer exhibit this preference directly and thus require an additional mechanism to explain the findings. Pickering *et al.* (2000) argue in favor of a probabilistic model, which combines traditional likelihood with a measure called *specificity* to explain these findings. The measure they derive is argued for on the grounds that it actually leads to a more optimal decision strategy than likelihood alone, under certain assumptions about the architecture of the human sentence processor (see also Chater, Crocker, & Pickering 1998)). In conclusion, we see further investigation and refinement of probabilistic models of human sentence processing as an enterprise, which we must seek to, and offers the best opportunity to, explain both the generally high standard of human linguistic performance, as well as specific pathological garden-path phenomena.

REFERENCES

Altmann, G. T. M., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 18*, 129–144.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioural and Brain Sciences, 14,* 471–517.

Brants, T. (1999a). Cascaded Markov Models, *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*, Bergen, Norway.

Brants, T. (1999b). *Tagging and parsing with Cascaded Markov Models—Automation of corpus annotation.* Vol. 6 of Saarbrücken Dissertations in Computational Linguistics and Language Technology, DFKI and Saarland University, Saarbrücken Germany.

Brants, T. (2000). TnT—A statistical part-of-speech tagger, *Proceedings of the 6th Conference on Applied Natural Language Processing,* Seattle, WA.

Brants, T., & Crocker, M. W. (2000). Probabilistic parsing and psychological plausibility, *Proceeding of the International Conference on Computational Linguistics (COLING 2000),* Saarbrücken, Germany.

Chater, N., Crocker, M. W., & Pickering, M. (1998). The rational analysis of inquiry: The case for parsing. In Chater & Oaksford (Eds), *Rational Analysis of Cognition,* (pp. 441–468). Oxford: Oxford University Press.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies, *Proceedings of the Annual Conference of the Association for Computational Linguistics,* Santa Cruz, California.

Corley, S., & Crocker, M. W. (2000). The modular statistical hypothesis: Exploring lexical category ambiguity. In M. W. Crocker, M. Pickering & C. Clifton (Eds.), *Architectures and mechanisms for language processing* (pp 135–160.) Cambridge: Cambridge University Press.

Crocker, M. W., & Corley, S. Modular architectures and statistical mechanisms: The case from lexical category disambiguation. In P. Merlo & S. Stevenson (Eds.), *The lexical basis of sentence processing,* New York, Benjamins, in press.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language, 27,* 429–446.

Ferreira, F., & Clifton Jr., C. (1986). The Independence of Syntactic Processing. *Journal of Memory and Language, 25,* 348–368.

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language, 26,* 505–526.

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

Juliano, C., & Tanenhaus, M. K. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society,* (pp. 593–598). Lawrence Erlbaum Associates.

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

Lapata, M., Keller, F., & Schulte im Walde, S. Verb frame frequency as a predictor of verb bias, submitted.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language, 32,* 692–715.

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9,* 157–201.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 10,* 676–703.

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

McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1998). Modelling the influence of thematic fit (and other constaints) in on-line sentence comprehension. *Journal of Memory and Language, 38,* 283–312.

Merlo, P., & Stevenson, S. (2000). Lexical syntax and parsing architecture. In M. W. Crocker, M. Pickering, & C. Clifton (Eds.) *Architectures and mechanisms for language processing,* (pp. 161–188). Cambridge: Cambridge University Press.

Pickering, M., Traxler, M., & Crocker, M. W. (2000). Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language, 43,* 447–475.

Rabiner, R. (1989). A tutorial on Hidden Markov Models and selected applications in??? recognition. *Proceedings of the IEEE, 77,* 257–285.

Ratnaparkhi, A. (1997). A linear observed time statistical parser based on maximum entropy. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Providence, Rhode Island.

Samuelsson, C. (1997). Extending n-gram tagging to word graphs. *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing,* Tzigov Chark, Bulgaria.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275,* 213–215.

Spivey-Knowlton, M. (1996). *Integration of visual and linguistic information: Human data and model simulations.* Unpublished doctoral disseration, University of Rochester, Rochester, N.Y.

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Hanna, J. E. (2000). Modelling discourse context effects: A multiple constraints approach. In M. W. Crocker, M. Pickering, & C. Clifton (Eds.) *Architectures and mechanisms for language processing* (pp. 90–118). Cambridge: Cambridge University Press.

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

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

Viterbi, A. (1967). Error bounds for convolution codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory, 13,* 260–269.