A Bayesian Model of Human Sentence Processing

Srini Narayanan
ICSI Berkeley
snarayan@icsi.berkeley.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

November 29, 2004

Abstract

1 Introduction

Language comprehension is a classic problem of reasoning under uncertainty. Language comes to us as a noisy, unsegmented, ambiguous mass of auditory waveforms or visual stimuli. Humans must somehow combine this input with other knowledge we have to come up with reasonable interpretations and actions. How might humans address this problem of decision-making under uncertainty? The best normative model we have for solving problems of this sort is probability theory, which offers a principled method with a coherent semantics for weighing and combining evidence. Whether this normative model is the correct descriptive model for all of human behavior has recently been the subject of much debate (Kahnema; Gigerenzer). While this debate is not resolved for all areas of human cognition and reasoning, the last decade or so had produced emerging consensus throughout the cognitive sciences that in some areas human cognition is likely to make use of probabilistic models. The seminal work of Anderson (1990) gave Bayesian underpinnings to cognitive models of memory, categorization, and causation, and recent Bayesian models of human cognition include work in human visual processing (Rao et al. 2001; Weiss & Fleet 2001), categorization (Tenenbaum, 2000; Tenenbaum & Griffiths, 2001b, 2001a), and the human understanding of causation (Rehder, 1999; Glymour & Cheng, 1998). Together, these ideas suggest that perhaps the process of human language comprehension is also best modeled as a process of probabilistic, Bayesian reasoning.

This idea that human processing of language draws on probabilistic models is hardly novel. (Schuchardt, 1885), in his arguments against the 19th century Neogrammarians, point out that key role of frequency in language production and language change. Schuchardt noted that word frequency is a good predictor of which words are phonologically weakened or 'lenited'. Words which are more frequent tend to be shorter and phonologically simplified; (Zipf, 1929) pointed out that this reduction of frequent forms also happened for frequent phones. (Jespersen, 1922) expanded Schuchart's idea from pure frequency to predictability or probability. Jespersen pointed out that the predictability of the word in its context, in addition to its raw frequency, must play a factor in the phonological form of the word.

These early intuitions about frequency and probability were all related to language production. Evidence for the role of frequency and probability specifically in language comprehension processing dates quite a bit later, from the mid 20th century. In the 1950's, for example, Davis Howes showed that word frequency plays a key role in comprehension in both the visual and auditory domains (Howes & Solomon, 1951; Howes, 1957). Throughout the second half of the 20th century, evidence amassed that high frequency words are accessed more quickly, they are accessed more easily, and they are accessed with less input signal than low-frequency words. This is a very robust effect, supported by tachistoscopic recognition Howes and Solomon (1951), naming (Forster & Chambers, 1973), lexical decision (Rubenstein, Garfield, & Millikan, 1970; Whaley, 1978; Balota & Chumbley, 1984), recognition accuracy and errors in noise (Howes, 1957; Savin, 1963), and gating (Grosjean, 1980).

The last two decades of behavioral research have extended these lexical results to other areas of psycholinguistics such as sentence processing. We know that many kinds of probabilistic knowledge play a role in the comprehension of sentences. One such factor is the probability of the different lexical categories of a word. For example the a priori probability that the ambiguous word *fires* is a noun, or alternatively a verb, plays a role in sentence comprehension, as does the probability that the word *selected* is a preterite or a participle (Burgess & Hollbach, 1988; Trueswell,

1996).This lexical category probability seems to be conditioned on context; thus for example the probability that the ambiguous word *that* will be determiner or a complementizer changes depending on the sentence context (Juliano & Tanenhaus, 1993). A very wide body of work has shown that a verb's subcategorization probability, for example the probability that a given verb is transitive or intransitive, plays a role in processing (Clifton, Jr., Frazier, & Connine, 1984; Ford, Bresnan, & Kaplan, 1982; Jennings, Randall, & TYler, 1997; MacDonald, 1994; Tanenhaus, Stowe, & Carlson, 1985; Trueswell, Tanenhaus, & Kello, 1993). Work in the last decade has extended this to simple semantic dependency probabilities such as the probability that a particular noun is the agent or patient of a particular verb (Trueswell, Tanenhaus, & Garnsey, 1994; McRae, Spivey-Knowlton, & Tanenhaus, 1998). We also know that local word-word relations such as the probability of a word given the previous or following words play a role in processing (MacDonald, 1993; McDonald, Shillcock, & Brew, 2001).

In summary, we know that many kinds of knowledge must interact probabilistically in the process of building an interpretation of a sentence. Unfortunately we still know very little about how this probabilistic process happens. We don't know how probabilistic aspects of linguistic knowledge are represented, we don't know how these probabilities are combined, we don't know how interpretations are selected, and we don't have a good understanding of the relationship between probability and behavioral measures like reading time.

Of course there has been quite a bit of research on the architecture of the human sentence processor over the last few decades, and this research has indeed touched on some of the probabilistic questions. But to a great extent, the field of sentence processing has asked other questions. Perhaps the largest area of focus in sentence processing has been on the debate surrounding the Modularity hypothesis of J.A. Fodor (Fodor, 1983). In J.A. Fodor's view, the human cognitive system is divided into 3 types of components: tranducers (sensory organs), input systems (vision, language), and central systems like executive and memory functions. Input systems are composed of modules which are domain-specific, informationally encapsulated, and localized in the brain. One extension of the modularity hypothesis has been to suggest that syntactic structural knowledge acts as a sort of sub-module. In this view, syntactic knowledge would lie in a module which is "informationally encapsulated" from the rest of linguistic knowledge. Furthermore, syntactic knowledge is assumed to be processed first, and so the earliest analysis of a sentence would only rely on syntactic knowledge. Real-world, lexical, and semantic constraints would come into play only later. Key to this line of research has been careful studies of the detailed time course of the activation of different knowledge sources, focusing on whether or not the use of syntactic knowledge precedes the use of semantic or lexical knowledge in the human sentence processor. (Ferreira & Clifton, Jr., 1986; Clifton, Jr. & Ferreira, 1987; Frazier & Fodor, 1978; Frazier, 1987; Frazier & Rayner, 1987; Frazier & Clifton, Jr., 1996). Arguing against this version of modularity has been a body of research focused on showing that a wide variety of constraints from the lexical, semantic, and extra-linguistic context plays an immediate role in processing (McRae *et al.*, 1998; MacDonald, 1993; MacDonald, Pearlmutter, & Seidenberg, 1994b; MacDonald, 1994; Spivey-Knowlton, Trueswell, & Tanenhaus, 1993; Spivey-Knowlton & Sedivy, 1995; Spivey & Tanenhaus, 1998; Tabossi, Spivey-Knowlton, McRae, & Tanenhaus, 1994; Trueswell & Tanenhaus, 1994; Trueswell *et al.*, 1994; Tanenhaus, Spivey-Knowlton, & Hanna, 2000; Tabor, Juliano, & Tanenhaus, 1997). Another key focus has been the role of memory, memory limitations, and locality in sentence processing. This area has focused on showing that memory limitations play a key role in explaining the complexity of processing certain sentences (Babyonyshev & Gibson, 1999; Gibson, 1998, 1990a, 1990b; Just & Carpenter, 1980; King & Just, 1991; Miyake, Carpenter, & Just, 1994).

Understanding the detailed time course of the use of different kinds of knowledge, and building a clear picture of the role that memory limitations, interference, and ' locality plays in processing are key aspects of the architecture of the human sentence processing mechanism. A complete understanding of sentence processing will need to somehow integrate these results into a single comprehensive model. But unfortunately most of these results don't say enough about the much more narrowly focused questions we posed above; how can we understand the role of probability in representing linguistic knowledge, combining evidence, and selecting interpretations.

One class of sentence processing models does address some of these questions about the role of probability. This is the framework generally called *constraint-based* or sometimes *constraint-based lexicalist* (MacDonald, Pearlmutter, & Seidenberg, 1994a; McRae *et al.*, 1998; Spivey-Knowlton *et al.*, 1993; Spivey-Knowlton & Sedivy, 1995; Seidenberg & MacDonald, 1999; Trueswell & Tanenhaus, 1994; Trueswell *et al.*, 1994; Kim, Srinivas, & Trueswell, 2002). Specific instantiations differ in various ways, but the shared intuition of the constraint-based models is that multiple interpretations of an ambiguous sentence are considered in parallel and that choice among these competing interpretations is made by integrating a large number of constraints over a wide variety of types of knowledge. Much research on the constraint-based models has focused on the time-course of constraint-access as part of the modularity debate dis-

cussed above, and hence is less relevant to our goals here. But particular instantiations of the constraint-based model have also led to specific claims about the representation, combination, selection, and behavioral (e.g., reading-time) implications of probabilistic knowledge.

There have been a number of computational implementations of the constraint-based framework, mainly neural-network models which take as input various frequency-based and contextual features, and combine these features via activation to settle on a particular interpretation (Burgess & Lund, 1994; Kim *et al.*, 2002; Spivey-Knowlton, 1996; Pearlmutter, Daugherty, MacDonald, & Seidenberg, 1994), but also including dynamical systems models (Tabor *et al.*, 1997). The most completely implemented of these models, and the one that makes the clearest claims about probabilistic integration and processing-time implications, is the competition-integration model of Spivey and colleagues (Spivey-Knowlton, 1996; McRae *et al.*, 1998), which uses a *normalized recurrence* algorithm for modeling constraint integration.

It's difficult to describe the model out of context, and therefore we will show the model as applied to a specific case of disambiguation in sentence processing. An understanding of this particular behavioral experiment will also prove useful as we compare the constraint satisfaction model with others that attempt to model this ambiguity. We will examine the Spivey model as applied by McRae *et al.* (1998) to the processing of sentences with the main-clause/reduced-relative clause (MC/RR) ambiguity. In these sentences, an initial sequence of words such as (1) is ambiguous. Continuations which illustrate the two possible parses, a subject noun phrase followed by a main verb, and a subject noun phrase postmodified by a reduced relative clause, are shown in (2) and (3):

(1) The witness examined

(2) The witness examined by the lawyer turned out to be unreliable.

(3) The witness examined the evidence.

These MC/RR ambiguities are known to cause processing difficulty, and have been used to test a wide variety of sentence processing models. In many cases, reduced relative clauses cause processing difficulty as measured by reading time increases at the disambiguating phrase. Many factors are known to play a role in the difficulty of these sentences. Trueswell *et al.* (1994) had shown that strong thematic constraints were able to ameliorate garden path effects in RR/MC ambiguities; subjects experienced more difficulty at the phrase "by the lawyer" in (4) than in (5). [1] The fact that *evidence* is a better *theme* than *agent* presumably provides evidence for the reduced-relative interpretation. As a result, the sentence processor may not settle on the main clause reading, reducing or eliminating the 'surprise' effect at the phrase *by the lawyer*.

(4) The witness examined by the lawyer turned out to be unreliable.

(5) The evidence examined by the lawyer turned out to be unreliable.

Various factors are known to play a role in processing such sentences, including the a priori probability that the verb (*examined*) is a preterite (simple past) versus participle, the general preference for main clause structures over reduced relative clause structures, the syntactic subcategorization bias of the the verb (*examined*), and the thematic fit of the subject head noun with the verb. Thematic fit is a measure of how likely a particular noun phrase is to appear as a particular thematic role for a verb. Thus *cop* is more likely to be the agent than the patient of *arrest*, i.e., a GOOD AGENT of *arrest*. *Crook* is more likely to be the patient, i.e., is a GOOD PATIENT of *arrest*.

McRae *et al.* (1998) had three goals. First, they wanted to confirm that thematic fit played a role in the disambiguation of MC/RR ambiguities. To this end, they need to show that Good Agent sentences like (6), in which the subject noun is biased toward an agent reading, produces a longer reading time at the phrase *the detective* than Good-Patient sentences like (7), in which the subject noun is biased toward a patient reading. Second, they showed that the competition model predicted these reading time differences.[2]

(6) The cop arrested by the detective was guilty of taking bribes.

[1] Although the original study by Ferreira and Clifton, Jr. (1986) had not found semantic effects, Trueswell *et al.* (1994) used a stronger manipulation of thematic constraint .

[2] They also had a third goal which we do not focus on, as it was part of an anti-modularity argument to show that thematic knowledge was used at the same time as syntactic knowledge.

(7) The crook arrested by the detective was guilty of taking bribes.

McRae *et al.* (1998) tested the competition model in both an off-line and on-line task. For the off-line task, they created 40 items, 20 with Good Agent subjects and 20 with Good Patient subjects. Subjects were given four iteratively longer sentence fragments for each item:

The crook arrested
The crook arrested by
The crook arrested by the
The crook arrested by the detective

Participants completed each fragment, and the proportion of main-clause and reduced relative completions was recorded.

In the on-line self-paced reading task, two complete sentence versions of each of the 40 items from the fragment task were created, one with a reduced relative clause, and one with an unreduced relative clause. The sentences were presented in a two-word moving window, as follows:

(8) The cop / arrested by / the detective / was guilty / of taking / bribes.

(9) The cop / who was / arrested by / the detective / was guilty / of taking / bribes.

Reading times were collected for three of these regions, the subject NP (*the cop*, the verb + preposition (*arrested by*) and the main verb group (*was guilty*).

Reading times for the unambiguous sentence in (9) were subtracted from the reading times for the ambiguous sentence in (8) to produce a delta reading time. Figure 1 shows this delta reading time for Good Agent and Good Patient sentences before and after the disambiguating region. As Figure 1 shows, the Good Agent sentences had a longer reading time at the disambiguating phrase *the detective* than the Good Patient did. This suggests that the Good Patient subjects biased the interpretation toward the reduced relative clause, eliminating this longer reading time.

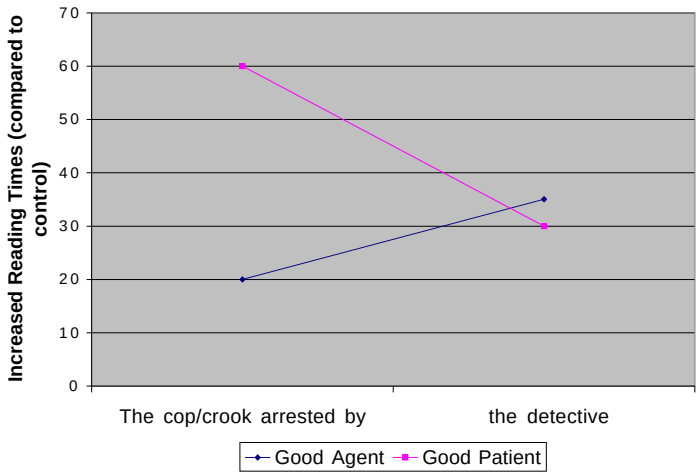

Figure 1: Self-paced reading times (from Figure 6 of McRae *et al.* (1998))

Both the completion and the reading time studies were then modeled by using the normalized recurrence component of the competition model to integrate various probabilistic constraints. A schematic of the model and the features is shown in Figure 2. There were six input features, four of which we will focus on:

syntactic bias toward main clauses: This feature represents the syntactic structural bias that favors main clauses and disfavors reduced relatives. Its value was set from the (Tabossi *et al.*, 1994) corpus counts of the percentage of times that the sequence 'NP verbed' was continued by a main clause, P(MC| "NP verbed") as opposed to a reduced relative P(RR| "NP verbed"),

participle versus preterite bias: This feature represents the preference of the main verb for a participle versus simple past reading. It was computed via the following equations:

$$\text{VTV(reduced)} = \frac{\log \text{Part} / \log \text{Base}}{\log \text{Part} / \log \text{Base} + \log \text{SP} / \log \text{Base}} \tag{10}$$

$$\text{VTV(main)} = \frac{\log \text{SP} / \log \text{Base}}{\log \text{Part} / \log \text{Base} + \log \text{SP} / \log \text{Base}} \tag{11}$$

***by* bias:** The bias the word *by* provides for a reduced relative interpretation. This was computed by counting in the Brown and Wall Street Journal corpora that the word *by* in each of the 40 verbs (in the "-ed" form) was followed by an agent and was in a passive construction, hence P(reduced relative|by,verbed).

thematic fit of initial NP: The fit of the subject as an agent of the verb, a number between 0 and 6 computed from role typicality ratings from a norming study.

The model uses a neural network to combine these constraints to support alternative interpretations in parallel. Each syntactic alternative (each 'parse') is represented by a single pre-built localist node in a network; thus the network models only the disambiguation process itself rather than the generation or construction of syntactic alternatives. The alternatives compete until one passes an activation threshold.

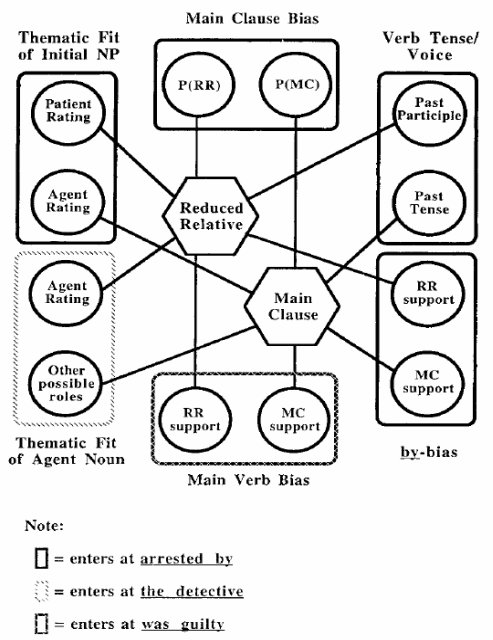

Figure 2: A schematic of the competition model, from McRae *et al.* (1998).

Each interpretation receives activation from the constraints which is then fed back to the constraint nodes within each cycle of competition. The algorithm first normalizes each pair of constraints. Let $C_{i,a}$ be the activation of the ith constraint node connected to the ath interpretation node. $C'_{i,a}$ will be the normalized activation; the activation of each constraint thus ranges from 0 to 1.

$$C'_{i,a} = \frac{C_{i,a}}{\sum_a C_{i,a}} \tag{12}$$

The activation I_a from the constraints to interpretation a is a weighted sum of the activations of the constraints, where w_i is the weight on constraint i (we will discuss below how the weights are set):

$$I_a = \sum_i w_i \times C'_{i,a} \tag{13}$$

Finally, the interpretations send positive feedback to the constraints:

$$C_{i,a} = C'_{i,a} + I_a \times w_i \times C'_{i,a} \tag{14}$$

As we saw above, each constraint i has a weight w_i. Each weight was set by searching through the space of all possible weight values for the set of values for all weights that made the model best fit the sentence completion data. These same weights were then used in the reading time data.

These three steps are iterated until one interpretation reaches criterion. The model predicts reading time in a similar way to other competition models like the construction-integration model of (Kintsch, 1988): reading time is modeled as a linear function of the number of cycles it takes an interpretation to reach criterion. Thus in general the closer two interpretations are in their activation values, the longer it will take for one of them to achieve a high enough activation to pass the threshold.

McRae *et al.* (1998) showed that the competition model predicted both the preferences expressed by the completion data, the reading times in by the self-paced reading task.

This competition-integration implementation of the constraint-based model fulfills a number of our criteria for a model which captures probabilistic effects in sentence processing. The model integrates a number of constraints, assigns each a probability value, combines the probabilistic constraints to predict a preference for ambiguous structures based on this probability value, and makes predictions about reading time based on the settling time of the competition between candidates.

While the constraint-based model is thus a good first step toward our goals, it still falls short in many ways. First, it is only a model of one aspect of the disambiguation process: choosing between ambiguous interpretations. The model thus doesn't have anything to say about how interpretations are constructed. Related to this problem is an unclarity with respect to the role of structural knowledge. The model includes a constraint preferring main-verbs to reduced-relative readings, based on a frequency difference in corpora. But no motivation is given for why this particular structural constraint is included and not any other. Certainly many other syntactic structures have large frequency differences, and are associated with different interpretations. Thus the model has no principled reason for choosing this constraint. In addition to problems with these structural aspects, the competition model uses constraint values that represent arbitrarily different probabilistic assumptions. Some are true probabilities, some are ratios of log probabilities, others are counts. Some are probabilities conditioned on the verb, some on the verb and the subject, some are not conditioned at all. The problem is not that there are different probabilities in the model, but rather that the model gives us no principle way to know which probabilities are included, and how they should be conditioned. Finally, the model makes use of various parameters (weights) that are used in combining the probabilistic constraints, but the model includes no component which tells us how to set these parameters.

In summary, the main problems with the constraint-satisfaction model have to do with structure; how structured interpretations are built probabilistically, how structural knowledge plays a role, what is the principled method for setting these probabilities of structure, and what the structure is of the algorithm for combining constraints. As it happens, there are alternative probabilistic models which focus on exactly these questions of structure. For example Jurafsky (1996) and Crocker and Brants (2000) both propose sentence processing models based on the intuitions of probabilistic grammars, which generally offer a principled foundation of probabilistic structure. Could these constitute an alternative instantiation of the constraint-based intuition?

In Jurafsky's model, a probabilistic parser keeps multiple interpretations of an ambiguous sentence, ranking each interpretation by its probability. The probability of an interpretation is computed by multiplying two probabilities: the stochastic context-free grammar (SCFG) 'prefix' probability of the currently-seen portion of the sentence, and the 'valence' (syntactic/semantic subcategorization) probability for each verb.

A stochastic context-free grammar, first proposed by Booth (1969), associates each rule in a context-free grammar with the conditional probability that the left-hand side expands to the right-hand side. For example, the following equations show the probability of two types of noun phrases, represented formally as two of the expansions of the nonterminal NP, computed from the Brown corpus:

[.42] NP $\rightarrow$ Det N
[.16] NP $\rightarrow$ Det Adj N

These rules tell us that the probability of expanded a noun phrase as a determiner followed by a noun is .42.

Jurafsky's model is on-line, using the left-corner probability algorithm of Jelinek and Lafferty (1991) and Stolcke (1995) to compute the SCFG probability for any initial substring (or 'prefix') of a sentence.

Subcategorization probabilities in the model are also computed from the Brown corpus. For example the verb *keep* has a probability of .81 of having two complements (*keep something in the fridge*) and a probability of .19 of having one complement (*keep something*). Jurafsky (1996) showed that this model could account for a number of psycholinguistic results on parse preferences. For example, the corpus-based subcategorization and SCFG probabilities for *keep* and other verbs like *discuss* correctly modeled the preferences for these verbs in the off-line forced-choice experiment of Ford *et al.* (1982).

While the model keeps multiple interpretations, it has only limited parallelism. Low probability parses are pruned via beam search, an algorithm for searching for a solution in a problem space that only looks at the best few candidates at a time. Because the model prunes interpretations (rather than keeping all possible interpretations around) means that occasionally the parse that was pruned will turn out to have been the correct parse. The model predicts extra reading time (the strong garden path effect) just in these cases, the correct parse had been pruned away and the rest of the sentence was no longer interpretable without reanalysis. Thus the Jurafsky (1996) model explains the misanalysis of garden path sentences like (15) and (16). In (15), the correct parse, in which raced is a reduced relative, is pruned. Thus when the parser arrives at *fell*, it is unable to integrate it into the parse, causing large reading time increases. In (16), the parse in which *houses* is a verb gets pruned, leaving only the nominal sense of *houses*. Later in the sentence, it becomes clear that the nominal sense of *houses* is incompatible with the sentence, again causing increased reading time.

(15) The horse raced past the barn fell.

(16) The complex houses married and single students and their families.

In these cases, the preference differences between the interpretations are modeled by combining the SCFG probability and subcategorization probability to compute a probability for each interpretation.

Crocker and Brants (2000) propose a similar probabilistic model of sentence processing that differs in using cascaded Markov models rather than SCFGs. Their *incremental cascaded Markov model* (ICMM) is based on the broad coverage statistical parsing techniques of Brants (1999). ICMM is a maximum-likelihood model, which combines stochastic context-free grammars with hidden Markov models, generalizing the HMM/SCFG hybrids of Moore, Appelt, Dowding, Gawron, and Moran (1995). The original non-incremental version of the model constructs a parse tree layer by layer, first at the preterminal (lexical category) nodes of the parse tree, then the next higher layer in the tree, and so on. In the incremental version of the model, information is propagated up the different layers of the model after reading each word. Each Markov model layer consists of a series of nodes corresponding to phrasal (syntactic) categories like NP or ADVP, with transitions corresponding to trigram probabilities of these categories. The output probabilities of each layer are structures whose probabilities are assigned by a stochastic context-free grammar. Figure 3 shows a part of the first Markov model layer for one sentence. Each Markov model layer acts as a probabilistic filter, in that only the highest probability non-terminal sequences are passed up from each layer to the next higher layer. The trigram transition probabilities and SCFG output probabilities are trained on a treebank.

The Crocker and Brants (2000) model accounts for various behavioral results on human parse preference, such as the Juliano and Tanenhaus (1993) studies on the disambiguation of *that*.

Both Crocker and Brants (2000) and Jurafsky (1996) have the advantages of a clean, well-defined probabilistic model which both explains how structures are built and how probabilities are assigned to them, both are incremental, showing how probability is computed word-by-word, and both offer a clear motivated probabilistic model of parse preference. In addition, the Jurafsky (1996) parser uses a parallel processing architecture which can capture the similarities between lexical and syntactic processing, and a probabilistic beam-search architecture which explains difficult garden-path sentences.

Unfortunately, neither of these models is sufficient to extend or replace the competition model as an explanation of human probabilistic processing. Unlike the Spivey model, neither the Jurafsky or Crocker and Brants models makes sufficient reading time predictions. The Crocker and Brants (2000) model is a model of preference, and as such does not make specific reading time predictions at all. The Jurafsky (1996) makes only very broad-grained reading-time predictions; it predicts extra reading time at difficult garden-path sentences, because the correct parse falls out of the parser's beam width. The Crocker and Brants (2000) model includes no verb valence model, and so it cannot model valence results of any kind, syntactic nor thematic, leaving it unable to model the wide variety of behavioral experiments showing the role of syntactic and semantic subcategorization, including the McRae *et al.* (1998) study.

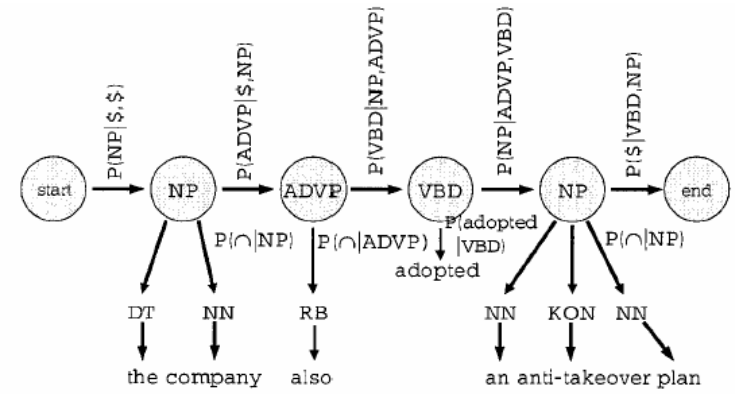

Figure 3: Part of the first layer Markov model for one sentence, from Crocker and Brants (2000). The letter *t* indicates the subtrees generated by the SCFG. Thus for example $P(t|NP)$ is the conditional probability of the subtree $NN \rightarrow$ *company* given the NP.

Furthermore, despite their probabilistic nature, neither the constraint satisfaction, Jurafsky (1996) or Crocker and Brants (2000) model a key class of behavioral studies on the probabilistic relation between individual words, often known as *word transition probabilities* or *word bigram probabilities*. McDonald *et al.* (2001) studied the effect of this probability on reading time by looking at eye fixations in subjects who were reading verb-noun pairs embedded in sentences. Subjects either read a sentence with a high transition probability verb-noun pair or a sentence with a low transition probability verb-noun pair. Other aspects of the sentence pairs, such as length and corpus frequency of the noun, neutral context, and sentence plausibility were all matched.:

high-probability: One way to **avoid confusion** is to make the changes during vacation.

low-probability: One way to **avoid discovery** is to make the changes during vacation.

McDonald *et al.* (2001) found that the duration of subjects' initial fixation on the target noun was shorter for the high-transition-probability verb-noun pairs. MacDonald (1993) reports on a similar earlier study using reading time.

The three probabilistic models also have a problem with modeling the behavioral results of Pickering, Traxler, and Crocker (2000), particularly since Pickering *et al.* (2000) argue that their results would cause problems for any frequency-based models. Pickering *et al.* (2000) looked at the disambiguation of the role of postverbal noun phrases in NP/S ambiguities. In the NP/S ambiguity, the postverbal noun phrase such as *his goals* in (17). can either be the direct object of the higher verb (an NP) (as in (18)) or be the subject of a sentential complement clause (an S), (as in (19)):

(17) The athlete realized $[_{NP}$ his goals] at the Olympics.

(18) The athlete realized $[_{NP}$ his goals] at the Olympics.

(19) The athlete realized $[_{S}[_{NP}$ his goals] were out of reach].

Previous research, as discussed earlier, suggests that verbs have a bias toward either an NP or S complement, and that this bias plays a role in processing. In the critical manipulation, Pickering *et al.* (2000) looked at sentences in

which the main verb was S-biased, like the verb *realize*. This means that at the point of reading the verb, human readers presumably expect the sentence to continue with a sentential complement. Pickering *et al.* (2000) then showed that if the postverbal noun phrase was an implausible direct object (like *her exercises* in (78) below) readers took longer to read the following words 'one day' than they did after a plausible direct object (like *her potential* in (77) below):

(20) The young athlete realized her potential one day might make her a word-class sprinter.

(21) The young athlete realized her exercises one day might make her a word-class sprinter.

In other words, *exercises* was anomalous only for one interpretation (the NP reading), but caused extra reading time. Pickering *et al.*'s result thus shows that a word which is anomalous only to the less-preferred interpretation causes a reading time increase.

The competition model has no way to account for this finding that decreasing the goodness of a less-preferred interpretation causes a reading time increase. Recall that in the competition (?), reading time increases are caused by competition between interpretations; the closer two interpretations are in preference, the longer it takes for a winner to settle out, and thus the longer the reading time. Thus the constraint-satisfaction model predicts that making the worse interpretation even worse should make the competition easier, hence *speeding up* the reading time, not slowing it down. In the model of Jurafsky (1996), reading time increases are caused by having to rebuild previously-pruned parses. But that cannot be the cause of the reading time increase on 'one day', since even if the direct-object parse becomes dispreferred enough to prune, it is the sentential complement parse that is continued in the rest of the sentence!

By contrast, Crocker and Brants (2000) note that their model can handle this result because they predict that it is the direct object reading, not the sentential complement reading, which is prefereed. Since their model does not compute valence probabilities of any kind, sentence preference is determined solely by the structure of the SCFG, and direct objects in general have a higher SCFG probability than sentential complements. Thus their model predicts that the probability of the direct object reading of (17) is actually higher than the probability of the sentential complement reading. Thus the extra reading time for the implausible direct object is explained by the fact that the direct object reading was the preferred one. The ability that lets the Crocker and Brants (2000) model handle this example, however, is its lack of valence probabilities. But this exact lack keeps their model from handling the extensive previous results showing the effect of verb bias (Clifton, Jr. *et al.*, 1984; Ford *et al.*, 1982; Jennings *et al.*, 1997; MacDonald, 1994; Tanenhaus *et al.*, 1985; Trueswell *et al.*, 1993). Thus in general, it is unlikely that this aspect of the Crocker and Brants (2000) model can be defended.

In summary, none of the three models we've looked at are sufficient for this data. The competition model has no way to build the two interpretations that it compares, no motivations for its weights, and no principled reason for using the specific conditional probabilities that it relies on. The Jurafsky (1996) model has an impoverished view of reading time and no clean way to combine information from multiple sources. Neither model can explain the Pickering *et al.* (2000) result. The Crocker and Brants (2000) model also offers no specific predictions about reading time, and is unable to model any effects of verb subcategorization or thematic preference. Although it would be easy to modify any of them, none of the models as described predict the word bigram probability result of McDonald *et al.* (2001).

In summary, no current models meet the criteria expressed above for modeling the role of probability in representing linguistic knowledge, combining evidence, selecting interpretations, and predicting behavioral variables like reading time.

Our goal in this paper is to attempt to build a model which meets these criteria. The fundamental insight of our model is the use of Graphical Models (specifically Bayes nets) in modeling the probabilistic, evidential nature of human sentence processing. Bayes nets are a type of model that can represent the causal relationship between different probabilistic knowledge sources, how they can be combined, and what we know about the independence of probabilities. In our Bayesian model of sentence processing, humans construct dynamic Bayes nets incrementally (on-line), while a sentence is being processed. Each Bayes net combines probabilistic knowledge of lexical, syntactic, and semantic knowledge on-line. Our proposal is thus that humans combine structure and evidence probabilistically, computing and incrementally re-computing the probability of each interpretation of an utterance as it is processed.

Like the Jurafsky (1996) model, this model is *on-line* and incremental; it assigns structure word by word as the sentence is read, changing structure as new information comes into the parser. Like most sentence processing models, our model is sensitive to various constraints, including syntactic structure, thematic biases, and lexical structure. Also like the Jurafsky (1996) model, our Bayesian model is probabilistic, incrementally computing the probability of each interpretation conditioned on the input words so far, and on lexical, grammatical and semantic constraints and knowledge. The most-preferred interpretation at any time is thus the one with the highest probability.

Our model differs from Jurafsky (1996) in two major ways. The first difference is that our model proposes a clean, principled way to combine structural knowledge and probabilistic knowledge: the Bayes Net or graphical model. Graphical models combine ideas from graph theory and probability theory to deal with two central issues facing large systems, *complexity* and *uncertainty*. Fundamental to these models is a notion of component composition - a complex system is built by composing simpler parts. Probability theory provides the glue that ensures that the combined system that comprises of simpler parts is consistent as a whole and interfaces as a whole to data. Graph theory provides a visual and intuitive interface as well as a formal data-structure that lends itself naturally to the design of efficient inference algorithms. Graphical models (Jordan 2003) present a common framework for many of the classical multivariate probabilistic systems - special cases of the general graphical model formalism include mixture models, factor analysis, hidden Markov models, Kalman filters and Ising models. The graphical model framework provides a way to view all of these systems as instances of a common underlying formalism.

The second difference is that our model makes fine-grained predictions about reading time. What kind of predictions could a probabilistic model make about reading time? One relation between probability and reading we have known for a long time. High frequency words are processed more quickly than low frequency words. As we mentioned earlier, this is a very robust result, whether from naming (Forster & Chambers, 1973), lexical decision (Rubenstein *et al.*, 1970; Whaley, 1978; Balota & Chumbley, 1984) or other metrics. We also know that predictability affects reading time. For example McDonald *et al.* (2001) modeled reading time results by showing that higher bigram predictability correlated with lower reading time. Indeed, the extra reading time associated with unexpected, anomalous, or unpredictable words has long been used as a methodological tool (for example with embedded anomalies). How can we cash out this relationship between frequency, predictability, probability, and reading time? In each of these cases, low probability (unpredictable) items are read more slowly while high probability items are read more quickly. Any probabilistic model predicts the probability of upcoming words. Thus any probabilistic model could model reading time by predicting that reading time is inversely proportional to the probability of upcoming words:

$$\text{reading time}(word) \propto \frac{1}{P(\text{word}|\text{context})} \tag{22}$$

Hale (2001) first noticed this fact about probabilistic parsers, and was the first to propose that this intuition could be probabilistically formalized in a probabilistic parser. His proposal is that the cognitive effort to integrate the next word into a parse is related to how surprising or unexpected that word is, and that this surprise be measured by the amount of information in the word. The information in a word can be measured information-theoretically as the negative log of its probability. Hale thus suggested that reading time was proportional to the negative log of the conditional probability of a word given the context.

Equation 22 gives us a key clue to making operational predictions about reading time from any probabilistic model. Like any incremental probabilistic model, our Bayesian model incrementally predicts the probability of all upcoming words. Our model integrates many sources of evidence (lexical, syntactic, semantic) into this probability computation. Thus the EXPECTATION component of our model predicts that the time to process an input (for example to read a word) is inversely proportional to the conditional probability of the word given the lexical, syntactic, and semantic evidence.

The second new way that the Bayesian model predicts increased reading time can also be viewed as a kind of expectation-based effect. Recall that our model is a parallel one, keeping multiple ranked interpretations. Our second prediction is that a demotion of this top interpretation causes extra reading time. For example, since probability is computed incrementally, an incoming word may make a previously dispreferred interpretation more likely, causing what had been the most-preferred interpretation to become second best. We predict that these demotions cause an increase in reading time. We refer to this prediction of our model as the ATTENTION principle, since it is based on our assumption that the comprehender places attentional focus on the best-ranked interpretation. Demotion of the interpretation in attentional focus causes increased reading time.

In summary, the goals of this paper are threefold. First, we introduce the idea of a Bayesian model for sentence processing. Our model suggests how the Bayes net could be used to represent probabilistic aspects of human knowledge of language, and how these probabilities are combined in computing the probability of an interpretation. Second, we propose a specific architecture for parsing, a probabilistic limited-parallel mechanism which makes specific predictions about the relationship between probability and reading time. Finally, we show that this model is able to account for behavioral results.

In the next section, Section 2, we lay out the model in detail, show exactly how the probabilities are assigned to

different parses of a word or sentence, how the Bayes Net is incrementally rebuilt as the sentence is processed word by word, and what the predictions are about behavior.

Sections 3, 4, and 5 then show the model's ability to handle behavioral data. Section 3 gives some motivating examples which show how probabilities of linguistic structure can be used to predict human preference in syntactic ambiguities. In section 4 and 5 we turn to the two reading time studies discussed above, McRae *et al.* (1998) and Pickering *et al.* (2000). It is important that our model be able to explain these behavior results for a number of reasons. First, the two studies cover the two most frequently-studied kinds of disambiguation: main clause/reduced relative and direct object/sentential complement. Second, no previous probabilistic model is able to account for the results of both these studies.

2 The Bayesian Model

The fundamental insight of our Bayesian model is to build multiple interpretations for the input, in parallel, compute the probability of each interpretation, and choose the interpretation with the maximum probability. Furthermore, this probability plays a role in reading time; words or structure which are unexpected (low probability) take longer to read.

In order to explicitly define our model and the probabilistic computations that it requires, we begin by introducing the basics of the use of Bayes rule for computing posterior probabilities.

We begin by considering an abstract form of the problem of choosing the highest probability interpretation. Imagine that we are given an input sentence s, and a set of potential interpretations i_1, i_2,...,i_n. Our task is therefore to compute the interpretation i^* that has the highest probability given the input sentence s. This can be expressed by the following formula:

$$i^* = \underset{i}{\operatorname{argmax}}\, P(i|s) \tag{23}$$

The function *argmax* returns the parameter which maximizes the value of its argument function. Thus equation (23) says that the best interpretation i^* is that particular interpretation i_j which has the property that its probability $P(i_j|s)$ is higher than the equivalent probability $P(i_x|s$ of any other interpretation i_x.

Equation (23) tells us that we could pick the maximum probability interpretation if we just knew how to compute $P(i|s)$ for each interpretation i and sentence s. One way to estimate a probability of an event is called the Maximum Likelihood Estimate: we simply count how many times the event occurs, and normalize by the count of all relevant events. So this suggests that we should compute $P(i|s)$ by asking 'out of every time that sentence s occurred in the past, how many times did it have interpretation i?'. While this is in fact mathematically correct, it cannot be the method that humans (or for that matter machines) use to compute the probability of interpretations. The reason, of course, is that language is creative and hence any given sentence is unlikely to have been ever uttered in the past, let alone enough times to estimate the probability of each of its multiple possible interpretations.

Since the human sentence processor cannot be computing interpretation probabilities by counting every time they occur in toto, we need a way to break down this probability computation down so that we are counting smaller pieces of an interpretation, each of which might have occurred often enough in the experience of a language user to be counted.

We propose that the human solution to this problem is based on two key ideas. The first key idea is *compositionality*: humans break down probabilities by making use of the compositional properties of language; a sentence is made up of words and syntactic structures. Generative linguistic theory provides us with good formal models of this kind of grammatical compositionality. Phrase-structure rule systems, or context-free grammar rules systems, such as X-bar phrase structure, provide a specific way to compose the structure of an entire sentence out of smaller pieces. These rules, as we will see in the next section, can be augmented with probabilities.

So we could compute probabilities for interpretations if we had a way to combine these probabilities for smaller structures into a single probability for an interpretation. The second key idea is thus a method for combining these probabilities together: the use of the *Bayes rule*. Bayesian reasoning is important because these modern models of linguistic structure are 'generative'. To simplify somewhat, a generative model is one that computes a string from some structure (for example a parse tree), rather than the other way around. For example, a phrase-structure grammar begins with a start-symbol S, and then using rules like $S \rightarrow NPVP$, expands this symbol, and then recursively expands the daughter symbols NP and VP to generate sentences. Because linguistic rules are generative, they are most naturally used to compute the probability that a particular interpretation 'generates' a particular sentence. In other words, it turns

out to be easier to compute $P(s|i)$ than $P(i|s)$. Luckily the Bayes Rule expresses a fixed mathematical relationship between $P(s|i)$ and $P(i|s)$, for any i and s, as follows:

$$P(i|s) = \frac{P(s|i)P(i)}{P(s)} \tag{24}$$

Bayes rule says that $P(i|s)$, the probability of an interpretation i given a sentence s, can itself be computed from three other factors. The first factor, $P(s|i)$, is called the *likelihood*. This is the probability of seeing the sentence if we were given that the interpretation was i; in other words, how likely the sentence s would be to occur if we knew the interpretation i was correct. The second factor, $P(i)$, is called the *prior*. This is the a priori probability of the interpretation before we had seen any new evidence. The denominator $P(s)$ is the probability of the sentence itself.

We can now use plug Bayes rule into (23)

$$i^* = \underset{i}{\operatorname{argmax}} \frac{P(s|i)P(i)}{P(s)} \tag{25}$$

We can simplify this equation a little. Consider the denominator term $P(s)$. This represents the probability of the sentence (sequence of words) s occurring. But equation (25) is asking 'For a given sentence s, what is the most probable interpretation?'. In other words, the string of words s is the same for each of the interpretations i. This means that we can just eliminate it from the equation, since multiplying each probability by a constant cannot change the ranking of probabilities that the argmax function is interpreting. Thus our final equation has only two components in the probability computation, the *likelihood* $P(s|i)$ and the *prior* $P(i|s)$:

$$i^* = \underset{i}{\operatorname{argmax}} \overbrace{P(s|i)}^{\text{likelihood}} \overbrace{P(i)}^{\text{prior}} \tag{26}$$

We will see the use of this combinations of likelihoods and priors in future sections as we introduce the various probabilistic estimators in our model (including the SCFG, valence probabilities, and N-gram probabilities).

The next section introduces the Bayes Net, the computational mechanism that we used for implementing the on-line probability computation that is central to our model, and the idea of conditional independence that underlies the Bayes Net. We then return to flesh out our various probability estimators.

2.1 Graphical Models as Probability Estimators

Our previous examples have dealt with complete sentences, and comparing probabilities for complete candidate parses. Indeed, statistical parsers were originally developed for text-processing purposes for which the entire sentence could be parsed at once. But human language processing is incremental, and so in our model parsing is done incrementally from left to right as each word is added to the input.

The advantage of a Bayesian approach to language processing is that it gives a model of what probability to assign to a particular source of evidence, and how these individual pieces of evidence combined in coming up with an overall interpretation that best fits the input. However, the sources of evidence update in an **incremental** fashion, as input comes in, and the posterior probabilities of different interpretations change. So we need a method to compute the incremental impact of new input.

We use Graphical Models (specifically Bayesian networks) for **on-line updates** of individual estimators; for example if we are estimating the probabilities of multiple possible interpretations of an ambiguous utterance, the network will allow us to compute the posterior probability of each interpretation as each piece of evidence arrives. In addition, the use of graphical models as a probabilistic estimator allows us to incorporate any kind of evidence; syntactic, semantic, discourse. This will allow us to capture the syntactic probabilities captured by graphical models like HMMs and SCFGs, while augmenting them with other probabilities, all in an on-line manner. Inference in graphical models relies on and directly exploits the structural aspects of the probability source. Technically, the inference procedures work by decomposing the overall *joint probability distributions* into a product of *conditional probability distributions*. This decomposition is based on exploiting the property of *conditional independence*.

2.1.1 Conditional Independence in Graphical Models

Independence is a powerful property because it allows us to reason about components in isolation. If A and B are probabilistic events, then event A is independent of event B iff $P(A) = P(A|B)$ or equivalently $P(A, B) = P(A) \times P(B)$. Note that is definition is symmetric. Thus is A and B are independent events, learning about the outcome of B does not impact the outcome probabilities of event A.

Unfortunately, most complex systems do not exhibit independence of components. For instance, a patient's symptoms are often not independent since they are manifestations of some specific disease. However it is often the case that specific conditioning variables (the interface variables) render independent the components of complex system. This generalization of the idea of independence is called *conditional independence*. Two components of a system exhibit conditional independence when the observation of a third aspect (set of variables) renders them independent. Thus the two components are independent given (conditioned on) the value of the third component. For example, knowing the disease (the conditioning variable) renders the symptoms independent.

It turns out that conditional independence does occur often in complex systems and leads to significant savings in representation and computation. Technically, if A, B, and C be events; A and B are conditionally independent given C iff $P(A|C) = P(A|B, C)$ or equivalently, $P(B|A, C) = P(B|C)$. Thus if A and B are conditionally independent, once we know the value of C, B (A) is independent (gives no additional information) about A (B).

Conditional independence assumptions are often made in many commonsense situations. Common examples include the assumption that symptoms are conditionally independent given a disease, and that the future and past are conditionally independent given the present (this is the famous Markov assumption inherent in markov models). A generalization of this notion of conditional independence is made in context free grammars. Here, we assume that the derivation of a non-terminal at a certain position in a parse tree (the outside probability) is independent of the derivation of the terminals dominated by the non-terminal in that position (the inside probability) given the identity and position of the non-terminal in question. This allows us to decouple the overall computation into top down and bottom up components which can be computed independently and combined for a specific node (position and value) in the parse tree. We will have more to say about this in the next section. In general, conditional independence is the key to reducing the representational and computational complexity in graphical models.

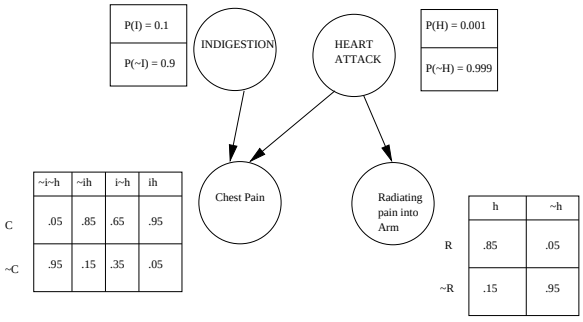

Figure 4: A simple Bayes Net for a diagnosing diseases. The two possible diseases, indigestion and heart attack can both cause chest pain, but only a heart attack can cause radiating arm pain. This is evident from the links in the network. Also knowing the disease is heart attack renders the symptoms conditionally independent.

Our model is based on a Directed Acyclic Graph (DAG) version of graphical models called a Bayes net. Figure 4 shows a simple illustrative Bayes net that models the relationship between a set of diseases and the symptoms they cause. The various nodes of interest (diseases, symptoms) are vertices in the graph. In general, a Bayes net consists of vertices which correspond to the variables of interest (such as possible non-terminals in a parse). When a node depends directly on another node, there is an edge between the appropriate vertices in the graph. Hence, in Figure 4, the vertex corresponding to the symptom (*chest pain*) has arrows coming from the diseases it's value *depends on* (in this case both *heart attack* and *indigestion*). If there is no direct dependence between two variables there is no edge between the appropriate vertices in the graph (there is no edge between *indigestion* and *radiating pain to the arm*).

The basic expressions in Bayes nets are statements about *conditional probabilities*. For example, $P(A|B)$ quantifies the belief in the proposition A given that the proposition B is known with absolute certainty. Graphical models use the principle of conditional independence as a basic representational primitive. Edges between nodes represent

direct influences between the variables.

The strengths of these influences are quantified by conditional probabilities; thus for each variable node A which can take values $a_1 \ldots a_n$, with parents $B_1, \ldots B_n$, there is an attached conditional probability table $p(A = a_1|B_1 = b_x, \ldots, B_n = b_z)$, $p(A = a_2|B_1 = b_x, \ldots, B_n = b_z)$, and so on. The conditional probability table (CPT) expresses the probabilities with which the variable A can take on its different values, given the values of the parent variables. In Figure 4, the CPTs for the individual disease nodes (HEART ATTACK and INDIGESTION) are not conditioned on any other variable (the disease nodes have no parents in the network in Figure 4). This unconditioned table represents the *prior* probability (absent any evidence) of the diseases. Thus Apriori, according to the network parameters, a heart attack is much less likely than indigestion (.001 to .1). The CPT values for the various symptoms are shown in the tables that quantify the influence on the specific symptom of the various joint assignments of values to the parent variables. In Figure 4, we see that for the variable CHEST PAIN, we have two parents (HEART ATTACK and INDIGESTION). The CPT for this variable has entries quantifying the likelihood of the symptom for all combinations of values of the parent. For instance, in the case that there is neither a heart attack nor indigestion, chest pain is relatively unlikely (.05), but when there is a heart attack and no indigestion, it is quite likely (.85), etc.

The distribution over all the variables (the joint distribution) can be compactly represented in Bayes nets. Figure 4, if we use the chain rule of probability, the joint probability of all the nodes is (by convention we will use lower case to indicate variables (so the variable i has two values, true and false):

$$P(i, h, r, c) = P(i) * P(h|i) * P(c|h, i) * P(r|h, i, c) \tag{27}$$

By using conditional independence relationships, we can rewrite this as

$$P(i, h, r, c) = P(i) * P(h) * P(c|h, i) * P(r|h) \tag{28}$$

where we were allowed to simplify the second term because h is independent of i and the last term because r is independent of i and c given its parent h.

We can see that the conditional independence relationships allow us to represent the joint more compactly. Here the savings are minimal, but in general, if we had n binary nodes, the full joint would require $O(2^n)$ space to represent, but the factored form would require $O(n2^k)$ space to represent, where k is the maximum fan-in of a node. And fewer parameters makes learning easier.

Bayes Nets can answer queries about any set of variables conditioned on any other set. The structure of the network reflects conditional independence relations between variables, which allow a decomposition of the joint distribution into a product of conditional distributions. The Bayes net thus allows us to break down the computation of the joint probability of all the evidence into many simpler computations. For example, in the example in Figure heart if there is no conditioning evidence, then indigestion is much more likely than a heart attack just based on the priors. Now suppose a patient comes in with chest pain. There are two possible causes for this: either he has had a heart attack, or he has indigestion. Which is more likely? We can use Bayes' rule to compute the posterior probability of each explanation (where f=false and t=true).The the chance of a heart attack conditioned on the symptom is

$$P(h = t|c = t) = \frac{P(h = t, c = t)}{P(c = t)} = \frac{\sum_{i,r} P(h = t, i, r, c = t)}{P(c = t)} = \frac{.001}{.111} = .01 \tag{29}$$

$$P(i = t|c = t) = \frac{P(i = t, c = t)}{P(c = t)} = \frac{\sum_{i,r} P(i = t, i, r, c = t)}{P(c = t)} = \frac{.06503}{.111} = .59 \tag{30}$$

The denominator for both calculations, $\sum_i \sum_r \sum_h P(C = t) = .111$ is the likelihood of the evidence.

In the case of a patient exhibiting chest pain ($c = t$), the network predicts an increased chance (compared to the Aprori value) of both heart attack and of indigestion, but not in the same proportions. Absent any confirming evidence of arm pain, the posterior probability of a heart attack ($P(h = t|c = t)$ is still one in a hundred (sixty times less likely than indigestion). Now, if new evidence comes in suggesting radiating pain to the arm ($r = t$), the picture changes and the posterior probability of a heart attack becomes much larger (.17) compared to the other diagnosis of indigestion. So as evidence comes in the posterior probability of different variables changes to reflect the effect of this new evidence. In general, the probability inference mechanism makes use of the the conditional independence

assumptions to simplify computing the distribution over the query nodes conditioned other nodes. These algorithms are called **belief propagation algorithms** (Pearl1988, Jensen 1995, Jordan 1999) and the exact computations are outside the scope of this paper. The reader is referred to the references above for a detailed treatment. We now turn to how the notion of how conditional independence informs the construction of our Bayes net model of sentence processing.

2.1.2 Exploiting Conditional Independence in Sentence Processing

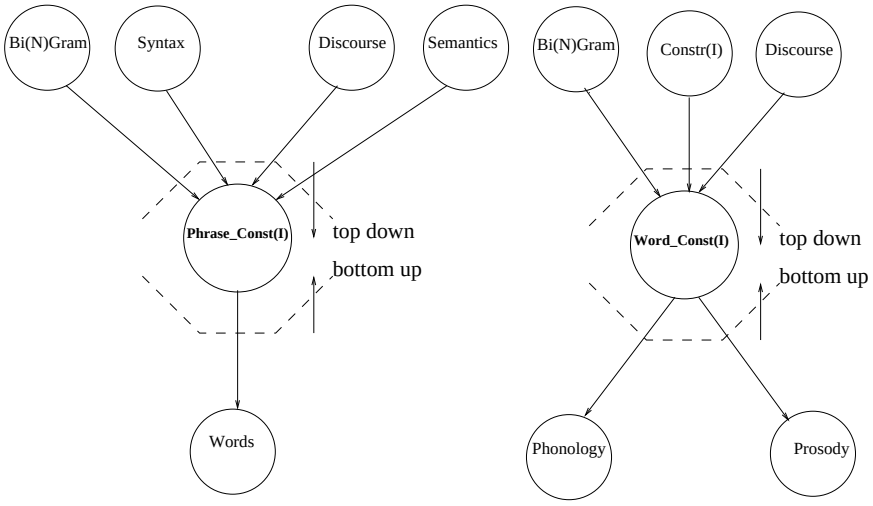

Figure 5: A Bayes net showing the structure of specific constructions at two different levels, the phrase level and the word level. In each level there are top down (syntactic, semantic, discourse, lexical, etc.) and bottom up sources (word, phonology, visual input, etc.). The network embodies the assumption that given a specific construction, the top-down and and bottom-up sources become conditionally independent. Uncertainty and structure go hand in hand all the way from the speech signal to discourse level constructions.

The crucial insight of our Bayes net model is to view specific interpretations as *latent variables* that render top-down (e^+) and bottom-up evidence (e^-) conditionally independent (d-separate them (Pearl, 1988)). We hypothesize that such a decomposition based on conditional independence exists at multiple levels all the way from the speech signal (or visual text perception) all the way upto phrasal and even discourse level constructions. Figure 5 shows two levels of the Bayes net structure which embodies this assumption. The figure on the left shows that a phrasal construction captures the correspondence between sets of words (bottom up) and a set of word associations (n-grams), syntax, discourse constraints and semantics (top down). The figure on the right shows that at a lower level of detail, a word construction may itself capture the correspondence between sets of phonological and prosodic features and higher level features including the phrasal construction that the word is part of. In interpretation, the top down constructional constrains provide expectations (of the next word or the next phonology) and the bottom-up constraints provide observational evidence of the recognized word (or phonology or prosody). Both top down and bottom up evidence are combined to arrive at an estimate of the overall support for a specific construction (phrasal or word level). We assume that there are similarly structured networks below the phonology level (where the bottom-up evidence may be features extracted from the speech signal) and above the phrase level (where the phrase level construction may provide bottom up evidence and discourse (and topic) level relations and information structure provide top-down evidence).

While our Bayes net based conceptual model supports integration of information across all these levels, the specific model implementation described in this paper is based on the phrasal construction level (the left side of Figure 5). In work described in this paper, we assume that lexical access (the right side of Figure 5) has already taken place. Syntactic, lexical, argument structure, and other contextual information acts as *prior* or *causal* support for an interpretation (like Main Clause or Reduced Relative), while bottom-up word strings other perceptual information acts as *likelihood*, *evidential*, or *diagnostic* support. Thus, it is a specific interpretation that captures explicit dependencies between syntactic, lexical and semantic sources. Technically, knowing the interpretation renders the various top down

sources (syntactic, semantic, discourse) *conditionally independent* of the bottom-up sources (words). Of course, as we will see in the next section, these sources (say syntax or semantic) are themselves structured and recursively decomposable (for instance the Context Free Grammar assumption for syntax) and the Bayes net formulation for the particular source directly reflects the various conditional independence assumptions made and takes advantage of the structure for inference. Our model is thus to have graphical models as individual probabilistic estimators for the various sources of support for a given construction. We then use a canonical technique for conjunctive source combination called NOISY-AND to come up with the overall estimate of the posterior probability for a specific construction.

To apply our model to on-line disambiguation, we assume that there are a set of interpretations ($(i_1, \ldots i_n) \in I$) that are consistent with the input data. At different stages of the input, we compute the posterior probabilities of the different interpretations given the top down and bottom-up evidence seen so far. As the input comes in, the posterior probabilities for the n interpretations are recomputed at different stages. Selection decisions thus depend on how an interpretation fits/explains the input (it's posterior value given the input). The interpretation that has the highest posterior at a given stage in the input is thus the best fit to the input at that stage. As the input comes in the fit of different interpretations shifts up or down by different amounts. We hypothesize that the reader is sensitive to certain types of unexpected shifts which results in enhanced reading times. Our Bayesian approach allows us to quantify and test this hypothesis for different types of reading time data. The next section outlines our model for computing the various probabilistic components that provide evidence for an interpretation and our Bayesian network implementation of the components.

We now describe how we compute the support for an interpretation from various sources (syntactic, lexical, thematic) at different stages in the input and then combine the individual source supports into an overall posterior probability of that interpretation at these input stages.

2.2 Individual Probability Estimators

We turn now to the details of computing the various probabilistic components of the Bayesian model: our goal is to arrive at an overall estimate $P(i|s)$, the posterior probability of an interpretation given a sentence (fragment). How should these probabilities be estimated? The computational linguistics literature abounds with methods for estimating these kinds of probabilities. One way to choose a method is to pick the simplest estimation algorithm that meets the constraints of psycholinguistic adequacy. We propose an algorithm consisting of only three relatively simple probabilistic components:

1. a *probabilistic word N-gram model*.
2. a *probabilistic syntactic* model
3. a *probabilistic verbal valence* model,

The next three sections will define each of these three components.

2.2.1 Word N-gram probabilities

The first component of our model captures the intuition that there is a probabilistic relation between adjacent words. Words are often very good probabilistic predictors of following words. We model this intuition with what is called *bigram probability*, *first-order Markov relation*, or sometimes *transition probability*: the conditional probability $P(w_i|w_{i-1})$ of a word w_i given a previous word w_{i-1}.

It is important to understand how transition probability differs from word frequency. A word can be rare, but be very predictable from the previous word. For example the word *havoc* is very rare (low frequency), but has a very high transition probability from the word *wreak*; thus $P(\text{havoc}|\text{wreak})$ is high.

Bigram probabilities can be easily computed from any corpus. The conditional probability of a particular target word w_i given a previous word w_{i-1} can be estimated from the counts of the number of times the two words occur together $Count(w_{i-1}w_i)$, divided by $Count(w_{i-1})$, the number of total times that the first word occurs:

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{31}$$

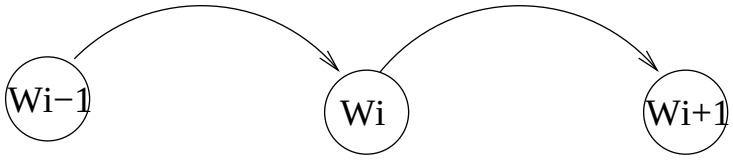

Figure 6: Bayes Nets for n-grams

2.2.2 Word N-gram probabilities in a Bayes Net

Probabilistic relations between adjacent words are modeled quite easily with a Bayes net (see Figure 6) that models a *bigram probability*: the conditional probability $P(w_i|w_{i-1})$ of a word w_i given a previous word w_{i-1}. Note that we can quite easily extend the graphical model to capture more complicated (trigram, n-gram) models.

2.2.3 A Probabilistic Syntactic model

There are a number of probabilistic models of syntactic structure (refs.). Of these, perhaps the earliest and simplest is the stochastic context free grammar. A stochastic context free grammar (SCFG) is a probabilistic version of the context free grammar (CFG) or phrase structure grammar. The non-stochastic CFG is widely used throughout linguistics and psycholinguistics as one of the mathematical skeleta which underlies generative linguistics models like principles and parameters, HPSG, and LFG (Sag-ref; Kaplan-ref).Phrase structure grammars reify assumptions about

word grouping and ordering that date as far back as the psychologist Wilhelm Wundt (1900), and were formalized by Chomsky (1956). We have chosen to use SCFGs to implement the structural portion of our model, because of their relative simplicity and wide-spread use. Our model could easily be adapted to other probabilistic models of syntactic structure. We have also chosen a very simple and theory-neutral version of SCFGs.

In each context-free production, an ordered list of words and phrasal symbols, appears the right of the arrow ($\rightarrow$) while to the left of the arrow is a single symbol expressing some cluster or generalization about these symbols. A CFG can assign a structure to an entire sentence, represented as a tree, by combining multiple rules. Figure 7 shows the tree representation of a derivation of the sentence 'The horse slept'. This derivation consists of 6 CFG rules:

S$\rightarrow$ NP VP
NP$\rightarrow$ DT NN
VP$\rightarrow$ VBD
DT$\rightarrow$ The
NN$\rightarrow$ horse
VBD$\rightarrow$ slept

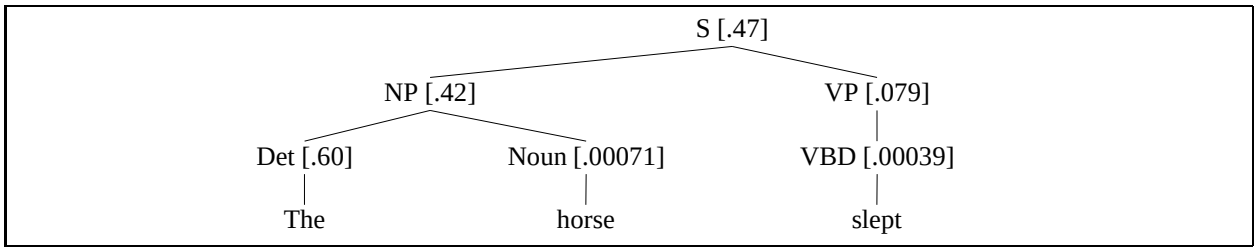

Figure 7: A parse tree for "The horse slept", with SCFG probabilities for the six rules.

A stochastic context-free grammar (SCFG) has the same phrase structure rules as a CFG. What an SCFG adds is that each context free rule is associated with a weight. This weight is the conditional probability of the right hand side of a rule given the left hand side, i.e. the probability of a particular expansion of a left-hand side. For example if the expansions of a verb phrase node *VP* consist of SCFG rules of the following form:

[.079] *VP* $\rightarrow$ *VBD*
[.18] *VP* $\rightarrow$ *VBD NP*
[.051] *VP* $\rightarrow$ *VBD NP PP*

then .079 is the probability that *VP* will expand to *V NP*, while .051 is the probability that *VP* will expand to *V NP PP*. These probabilities are all conditional on *VP*, which means that the probabilities of all the expansions of a given nonterminal sum to one. These probabilities can be computed from a *treebank* (a parsed corpus) by counting the number of times each kind of rule expansion occurs.

The probability for an entire parse tree T and the surface sentence it produces S is found by multiplying together the probabilities for the CFG rules used to expand each node in the tree. Thus the probability of the entire structure shown in Figure 7, that is to say the probability of the parse tree together with the string of words, is derived as follows:

$$\begin{aligned} P(T,S) &= P(\text{NP, VP, Det, Noun, VBN}|S, the, horse) \\ &= P(S \rightarrow \text{NPVP}, \text{NP} \rightarrow \text{DetNoun}, \text{VP} \rightarrow \text{VBD}, \text{Det} \rightarrow the, N \rightarrow horse)\text{VBD} \rightarrow slept \\ &= P(S \rightarrow \text{NPVP}) \times P(\text{NP} \rightarrow \text{DetNoun}) \times P(\text{VP} \rightarrow \text{VBD}) \times P(\text{Det} \rightarrow the) \times P(N \rightarrow horse) \times P(\text{VBD} \rightarrow slept) \\ &= .47 * .42 * .079 * .60 * .00071 * .00039 = .0000000025 \end{aligned} \quad (32)$$

In other words, the probability of the sentence and the tree is the product of the probabilities of each of the six rules. More generally, for each node n in the parse tree *Tree*, let *rule_expansion(n)* represent the rule which expands that node. Then:

$$P(\text{Tree}, S) = \prod_{n \in \text{Tree}} p(\text{rule\_expansion}(n)|n)$$

FIX: REPLACE WITH THE EVIDENCE EXAMINED. Figure 8 shows the SCFG parse tree for the sentence *The horse raced past the barn fell*, which will play a role in our later descriptions of our sentence processing algorithm.

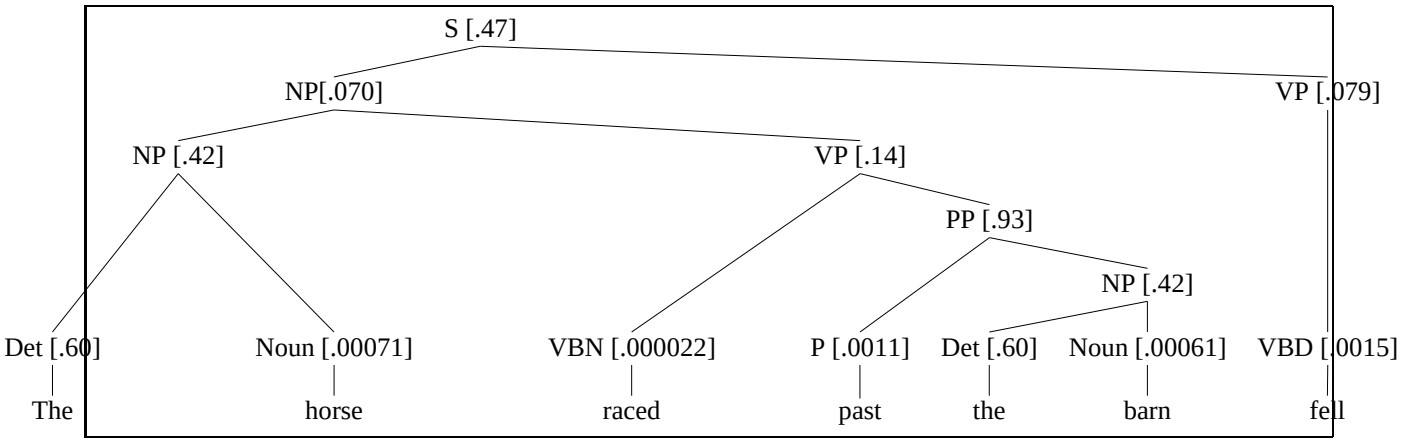

Figure 8: A parse tree for "The horse raced past the barn fell", with SCFG probabilities for the rules. These probabilities were drawn from the Penn Treebank annotation of the Brown corpus, except for the rule *VBN* $\rightarrow$ *raced*, which didn't occur in the Brown corpus, and was estimated using the web.

2.2.4 Processing SCFG probabilities with the Bayes net

We begin with a description of how the Bayes net computes the SCFG portion of our probabilistic model in an **incremental** fashion. In this paper, we will consider only Bayes nets for previously generated partial parse tree structures. We assume the presence of a chart or some mechanism to dynamically generate the partial parse trees for the input, given a grammar. Given a parse structure, we generate the appropriate Bayes net and compute the posterior probabilities for the competing interpretations. It turns out that we can set up a relatively straightforward correspondence between the computation of SCFG probabilities by a probabilistic parsing algorithm (as described in the previous section) and the 'belief propagation' algorithm of Bayes nets.[3]

The correspondence is as follows:

[3]More technically, for those who are interested, the Inside/Outside algorithm applied to a fixed parse tree structure is obtained exactly by casting parsing as a special instance of belief propagation (Narayanan 2004).

- The parse tree is interpreted as a Bayes net.
- Non-terminal nodes in the parse tree correspond to nodes in the Bayes net, the range of the variables being the non-terminal alphabet.
- The grammar rules define the conditional probabilities linking parent and child nodes.
- The S nonterminal at the root, as well as the terminals at the leaves represent conditioning evidence to the network.
- Conditioning on this evidence produces exactly the conditional probabilities for each nonterminal node in the parse tree and the joint probability distribution of the parse.[4]

Consider the partial parse state after the input "The horse raced". Figure 9 shows the partial SCFG parses for the main clause, MC and reduced relative, RR interpretations for this input (recall the definition of the main clause and reduced relative interpretations of 'The horse raced past the barn…' in Section 2.2.3.)

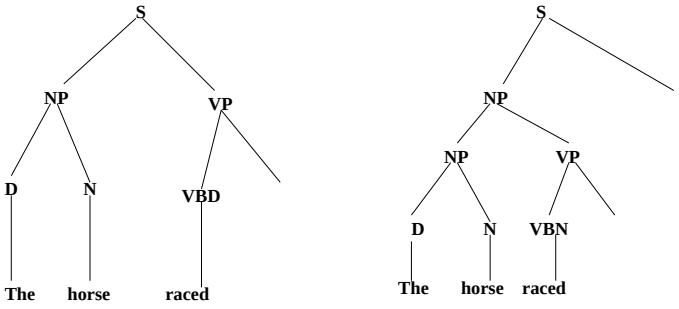

Figure 9: MC and RR SCFG parse states (parses in chart) for the input 'The horse raced…'.

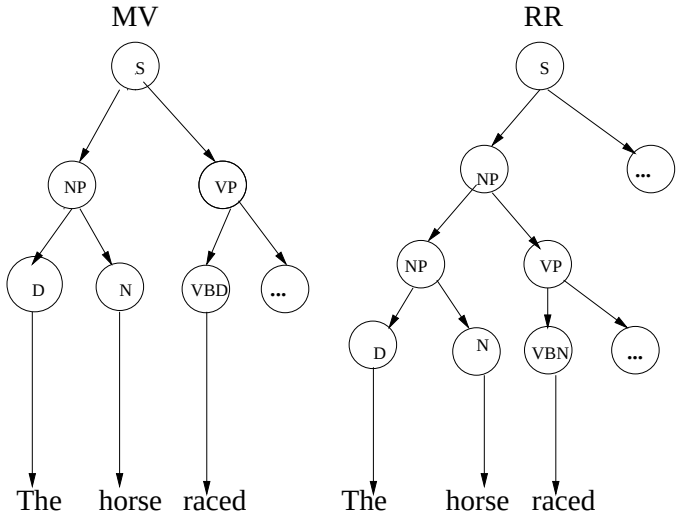

Figure 10: Pieces of Bayes networks corresponding to two SCFG parses for the prefix 'The horse raced …'. The … label on specific nodes in the Bayes net indicates sums over all continuations of the partial parse state.

[4]One complication is that the the conditional distribution in a parse tree $P(Y,Z|X)$ is not the product distribution $P(Y|X)P(Z|X)$ (it is the conjunctive distribution). However, it is possible to generalize the belief propagation equations to admit conjunctive distributions $P(Y,Z|X)$ and $P(X,V|U)$. The diagnostic (inside) support becomes $\lambda(x) = \sum_{y,z} \lambda(y)\lambda(z)P(y,z|x)$ and the causal support becomes $\pi(x) = \beta \sum_{u,v} \pi(u)\lambda(v)P(x,v|u)$ (details can be found in Appendix A).

Figure 10 shows the equivalent Bayes net for the parse tree in Figure 9. The probability of the MC parse can be computed by belief propagation on the Bayes net in Figure 10. Note the conditional independence statements reflect the context free assertion made by SCFG grammars. At this point in the input, the network expresses the active structures after seeing the word *the horse raced*[5].

At this stage of the input, the network is thus computing the following probabilities:

The MC and RR parse likelihoods, given the input and the conditional independence assertions embodied in the Bayes net is

$$\begin{aligned} P(T,S)^t_{MC} &= P(\text{NP, VP, Det, Noun, VBD, } \textstyle\sum(\ldots)|S, the, horse, raced) \\ &= P(S \rightarrow \text{NPVP, NP} \rightarrow \text{DetNoun, VP} \rightarrow \text{VBD}\textstyle\sum(\ldots), \text{Det} \rightarrow the, N \rightarrow horse\text{VBD} \rightarrow raced) \\ &= P(S \rightarrow \text{NPVP}) \times P(\text{NP} \rightarrow \text{DetNoun}) \times P(\text{VP} \rightarrow \text{VBD}\textstyle\sum(\ldots)) \times P(\text{Det} \rightarrow the) \times P(N \rightarrow horse) \\ &\quad \times P(\text{VBD} \rightarrow raced) \end{aligned} \tag{33}$$

$$\begin{aligned} P(T,S)^t_{RR} &= P(\text{NP}, \textstyle\sum(\ldots), \text{NP, VP, Det, Noun, VBN, } \textstyle\sum(\ldots)|S, the, horse, raced) \\ &= P(S \rightarrow \text{NP}\textstyle\sum(\ldots), \text{NP} \rightarrow \text{NPVP, NP} \rightarrow \text{DetNoun, VP} \rightarrow \text{VBN}\textstyle\sum(\ldots), \text{Det} \rightarrow the, N \rightarrow horse\text{VBN} \rightarrow raced) \\ &= P(S \rightarrow \text{NP}\textstyle\sum(\ldots)) \times P(\text{NP} \rightarrow \text{NPVP}) \times P(\text{NP} \rightarrow \text{DetNoun}) \times P(\text{VP} \rightarrow \text{VBN}\textstyle\sum(\ldots)) \times P(\text{Det} \rightarrow the) \\ &\quad \times P(N \rightarrow horse) \times P(\text{VBN} \rightarrow raced) \end{aligned} \tag{34}$$

Figure 11 shows the equivalent Bayes net for the parse tree obtained at a future stage $(t+k)$ after the input *the horse raced past the barn.*

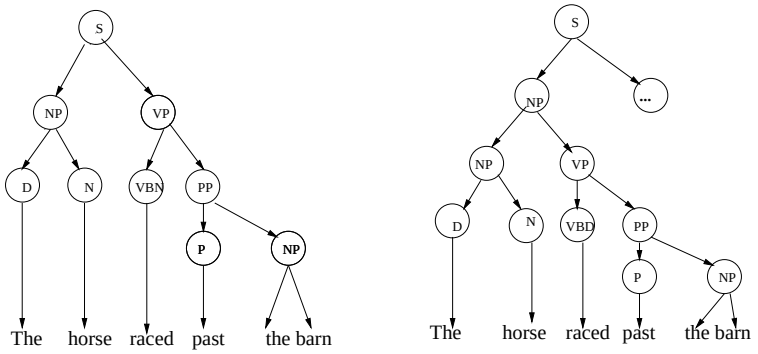

Figure 11: Pieces of Bayes networks corresponding to two SCFG parses for the prefix 'The horse raced past the barn…'. The …label on specific nodes in the Bayes net indicates sums over all continuations of the partial parse state.

The MC and RR parse likelihoods, given the input and the conditional independence assertions embodied in the Bayes net is

$$P(T,S)^{t+k}_{MC} = P(\text{NP, VP, Det, Noun, VBD}\ldots|S, the, horse, raced, past, the, barn)$$

[5]Note that the use of the … labeled nodes are technically the result of summing over all possible completions of the SCFG structure starting with the specific prefix non terminal. This part of the model is similar to the Jurafsky (1996) model and uses well known algorithms to compute the prefix probability (Jelinek and Lafferty 1991; Stolcke 1995) for a **given** SCFG grammar.

$$
\begin{aligned}
&= P(S \rightarrow \text{NPVP}, \text{NP} \rightarrow \text{DetNoun}, \text{VP} \rightarrow \text{VBDPP}, \text{Det} \rightarrow the, N \rightarrow horse, \text{VBD} \rightarrow raced, \\
&\quad \text{PP} \rightarrow \text{PNP}, \text{P} \rightarrow past, \text{NP} \rightarrow \text{DetN}, \text{Det} \rightarrow the, \text{N} \rightarrow barn) \\
&= P(S \rightarrow \text{NPVP}) \times P(\text{NP} \rightarrow \text{DetNoun}) \times P(\text{VP} \rightarrow \text{VBD}\textstyle\sum(\ldots)) \times P(\text{Det} \rightarrow the) \times P(N \rightarrow horse) \times \\
&\quad P(\text{VBD} \rightarrow raced) \times P(\text{PP} \rightarrow \text{PNP}) \times P(\text{P} \rightarrow past) \times P(\text{NP} \rightarrow \text{DetN}), \times P(\text{Det} \rightarrow the), \\
&\quad \times P(\text{N} \rightarrow barn) \qquad (35)
\end{aligned}
$$

$$
\begin{aligned}
P(T,S)_{RR}^{t+k} &= P(\text{NP}, \text{VP}, \text{Det}, \text{Noun}, \text{VBN}, \textstyle\sum(\ldots) | S, the, horse, raced, past, the, barn) \\
&= P(S \rightarrow \text{NP}\ldots, \text{NP} \rightarrow \text{NPVP}, \text{NP} \rightarrow \text{DetNoun}, \text{VP} \rightarrow \text{VBNPP}, \text{Det} \rightarrow the, N \rightarrow horse \text{VBN} \rightarrow raced, \\
&\quad \text{PP} \rightarrow \text{PNP}, \text{P} \rightarrow past, \text{NP} \rightarrow \text{DetN}, \text{Det} \rightarrow the, \text{N} \rightarrow barn) \\
&= P(S \rightarrow \text{NP}\textstyle\sum(\ldots)) \times P(\text{NP} \rightarrow \text{NPVP}) \times P(\text{NP} \rightarrow \text{DetNoun}) \times P(\text{VP} \rightarrow \text{VBNPP}) \times P(\text{Det} \rightarrow the) \\
&\quad \times P(N \rightarrow horse) \times P(\text{VBN} \rightarrow raced) \times P(\text{PP} \rightarrow \text{PNP}) \times P(\text{P} \rightarrow past) \times P(\text{NP} \rightarrow \text{DetN}), \\
&\quad \times P(\text{Det} \rightarrow the), \times P(\text{N} \rightarrow barn) \qquad (36)
\end{aligned}
$$

Of course at the next stage in the input, the sentence ending marker after the word "fell." leaves only one interpretation, namely the RR interpretation.

2.2.5 Lexical Valence Probabilities

The syntactic part of our model, SCFG, was used to capture structural facts about grammatical knowledge. The third part, a model of *probabilistic verbal valence*, is designed to capture *valence* knowledge, the biases and expectations that a predicate (such as a verb) has for its arguments.

In most proposals for the lexical representation of verbal semantics, the verb has expectations for particular *thematic roles*. Some verbs expect roles like agent and theme, other expect propositions, and so on, Our model expresses the probability that potential arguments play particular thematic roles in the verb. This thematic role probability expresses the probabilistic dependency relation that a verb has in assigning a particular thematic role to a particular argument. This probability is conditioned on the head words of the argument and on their syntactic position.

For example, the verb *elected* may have a preference to assign an agent role to the subject noun phrase *they*, but a patient role to the object noun phrase *them*. Or the verb *open* may prefer to assign the thematic role *agent* to the subject 'The window-cleaner' but the thematic role *theme* to the subject 'The window'.

Since verbs may have more than one argument, this verb-argument expectation can be expressed as the expectation of a verb for a set of arguments and their thematic roles. For example equation 37 expresses the probability of the verb *take* assigning the Agent semantic role to its subject NP *they* given all the other arguments in the sentence.

$$
P(\text{subject=Agent} | \text{verb=take}, \text{subject=they}, \text{object=books}, \text{toPPcomp=to the library}) \qquad (37)
$$

While these probabilities can in principle be computed from corpora, current corpora do not seem to be big enough to contain such specific counts. For example, the 100 million word British National Corpus contains no instances of the verb "fired" with a subject NP whose head noun is "employer" or an object NP whose head noun is "employee".

This suggests that people may also not be storing this exact probability. They may instead be approximating it in various ways. One way to approximate this probability is to assume that they are stored over semantic clusters of words rather than individual words. Thus eventually methods such as clustering or other uses of semantic features can be used to generalize corpus counts. Another way to approximate this probability is to assume that the probabilities of the individual arguments are independent. We would then be computing expectations separately for each argument. For example for the thematic role of the subject of the verb *elect*, we would compute the following conditional probabilities:

$$
\begin{aligned}
&P(\text{Agent} | \text{verb=elect}, \text{subject="they"}) \\
&P(\text{Theme} | \text{verb=elect}, \text{subject="they"})
\end{aligned}
$$

This thematic fit probability would be computed separately for every potential argument, not just the subject. Thus the preposition phrase *by the cop* in the sentence *The crook was arrested by the cop,* plays the agent role for

arrest. Thus the probability expressed by Equation 38 will presumably be higher than the probability expressed by Equation 39:

$$P(\text{Agent}|\text{verb=arrest, byPP="the cop"}) \tag{38}$$

$$P(\text{Theme}|\text{verb=arrest, byPP="the cop"}) \tag{39}$$

Even these less complex independent probabilities for each argument require counts that are too rare to find in a corpus. Eventually these can be computed via clustered models or by computing probabilities over semantic features rather than words. In the current study we relied on norming studies for each of the behavioral experiments we are modeling. In one experiment, we will model data from McRae *et al.* (1998), using their norming study counts. In their study, the typicality of the noun as a filler of the agent versus theme role was determined by having 36 subjects complete a rating task, answering questions like the following:

(40) How common is it for a crook to **arrest** someone?

(41) How common is it for a crook to be **arrested by** someone?

Their subjects judged role filler typicality on a 1-7 Likert; 1 corresponded to a very uncommon event, and 7 to a very common event. We converted these numbers to probabilities by norming them (dividing the value by 7 to get a probability value between 0 and 1).

Our second set of experiments model data from Pickering *et al.* (2000), using probability parameters taken directly from the Pickering study itself. They asked subjects to complete sentence fragments like "The young athlete realized" or "The young athlete realized her", and counted the number of times that the completions were syntactic direct objects or semantic THEMES ("The young athlete realized her goals") versus the number of times that the completions were syntactic sentential complements or semantic PROPOSITIONS ("The young athlete realized her exercises weren't working"). The result was a set of probabilities of the THEME or PROPOSITION argument. given the verb, the initial NP, and the word "her". Because in these cases THEMES correspond to DOs, and PROPOSITIONS correspond to SCs, Pickering *et al.* (2000) referred to these as SC versus DO probabilities, as follows:

$(P(SC)|V = realized)$
$(P(DO)|V = realized)$
$(P(SC)|VP = realized, [_{NP}her \ldots], InitialNP = The, young, athlete)$
$(P(DO)|VP = realized, [_{NP}her \ldots], InitialNP = The, young, athlete)$

Our model thus includes syntactic subcategorization probabilities as well as the thematic subcategorization probabilities we have been discussing. We believe it is likely that it will be possible to modify the model to rely solely on thematic probabilities; that is, thematic probabilities may obviate the need for syntactic subcategorization probabilities. However we save this research to be addressed in future work.

In general we believe our estimates of valence probabilities are quite rough. As was true with our SCFG probabilities, our goal is not to test this particular model of valence representation, but to show how an approximate instance of this knowledge type can be incorporated into our Bayesian model.

2.2.6 Processing valence probabilities with the Bayes net

Figure 12 outlines the basic structure of the Bayes net for processing the valence probabilities. In our model, we compute the support for an interpretation given a verb and all its subcategorization information (syntactic arguments (Syncat) and their fit to verb specific thematic/frame roles (Role)). In addition, our model makes the following conditional independence assumptions.

1. The identity of the verb determines its argument structure bias. This is shown in Figure 12 as the node labeled $Frame$ depends only on the identity of the verb.

2. The thematic fit of an input phrase (for the different roles (theme, agent)) depends only on the identity of the verb and the potential argument fillers in the input (labeled as Arg1, Arg2, ... Argn). The conditional independence assumption made here states that dependencies between a verb and it's syntactic arguments is captured by the various semantic roles. This leads to the structure in the second row of the Bayes net in Figure 12.

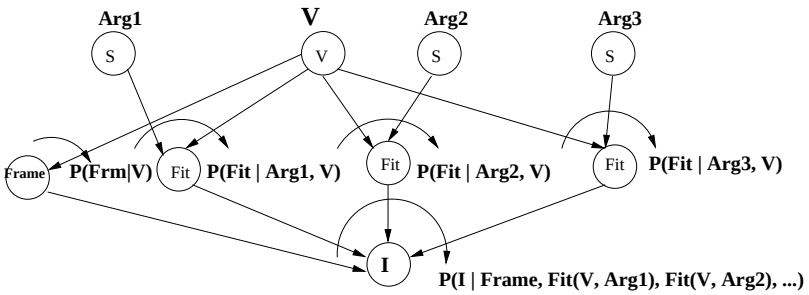

Figure 12: Bayes Nets for valence Probabilities. Given a verb (predicate), the probability of frame (thematic) roles fitting specific argument fillers is computed along with the subcategorization bias (frame) for the verb (predicate).

3. Specific interpretations constrain the various thematic fits for a given verb frame. For instance, the reduced relative interpretation (I) requires that the subject (Arg1) be the theme (Fit) for a transitive verb. Thus a specfic interpretation (I=reduced relative) captures the dependencies between the verb frame (frame=transitive) and the various thematic role bindings (such as Arg1:subj=theme). This is the conditional independence assumption that leads to the third row in the Bayes net of Figure 12.

   The second row in Equation 42 embodies the three conditional independence assumptions stated above. made in this computation.

$$
\begin{aligned}
P(I\_val) &= P(I|V, arg_1, arg_2, \ldots arg_n, them\_fit(V, R_i, arg_j) : \forall i \in Roles(frame)) \\
&= P(I|frame(V), them\_fit(R_i|arg_j, V) : \forall i \in Roles(frame), j \in Syncat(frame)
\end{aligned}
\tag{42}
$$

The first two conditional independence assumptions result in the middle layer of Figure 12 and the final dependence is captured by the bottom layer in Figure 12. We now turn to how this Bayes network is used to model the lexical/thematic dependencies for the studies reported in this paper.

Figure 13 and Figure 14 show the Bayes net for the lexical and valence probability computations for the McRae *et al.* (1998) data. Figure 13 (top row) shows the structure and probabilities encoded. In general we quantify the semantic fit (Agent or Theme) based on the identity of the verb and the syntactic category argument (subject, object, etc.) We also quantify the argument structure (transitive or intransitive) preference (bias) based on the identity of the verb. Here $Arg1$ is the Subject NP and $Arg2$ is the by Prepositional Phrase $byPP$. The networks at the bottom row show the MC (left network) and RR (right network) for the input "The witness examined by the lawyer turned out to be unreliable" at different stages of the input.

In all these cases, the node labeled V (the root node of the semantic Bayes net) represents a variable that ranges over the set of verbs. For a particular verb (like *arrest*), this node would set to a particular value (V = arrest). The node labeled $Frame$ has values [transitive, intransitive]. The conditional probability values quantify the probability that the $Frame$ node has a specific value (trans, intrans) **given** the identity of the verb (such as arrest, examine race). So if the domain of interest were restricted to the three verbs (arrest, examine and race) , the table entries for the $Frame$ node would be as follows in the second column of Table 1 below. The various thematic fit conditional probability distributions are shown in the third column of the table.

As input comes in, more of the lexical/thematic network gets instantiated (more $argi$ nodes have values) and the fit of the new potential argument nodes to frame/thematic roles can be computed. As in other models (Gildea and Jurafsky 2001), we assume an enumerated set of possible thematic roles sometimes aggregating over these roles with the OTHER value for the fit (as shown in Figure 14). As in the case of Bayes net model of syntax (SCFG), input coming in allows for the re-estimation of posterior probabilities for the different interpretations.

Table 1 and Table 2 show examples of parameters encoded in the lexical Bayes net in Figure 13. Table 1 pertains specifically to the network in Figure 13. In our model, different interpretations impose constraints on the network. We evaluate the constrained networks to compute the total posterior for a particular interpretation. The bottom left and bottom right networks in Figure 13 specify the constraints for the MC and RR interpretations respectively. For

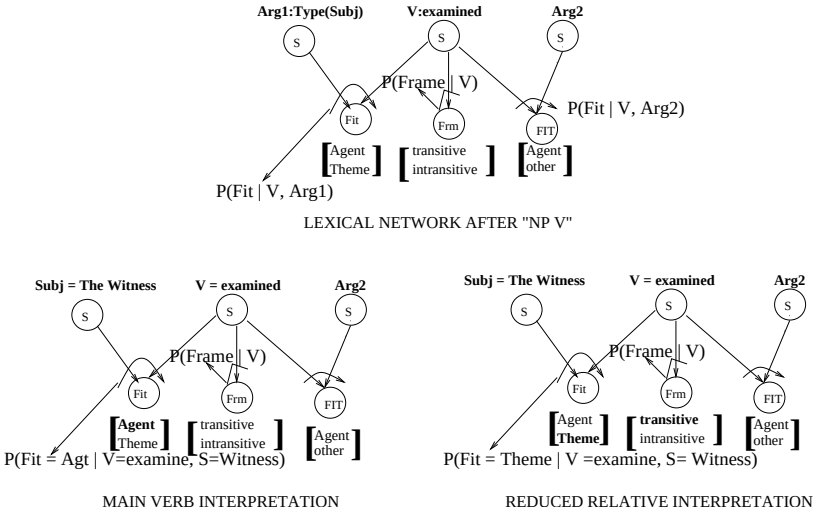

Figure 13: The Lexical/Thematic Bayes net for valence for the Mcrae MC/RR example. The bottom left shows the network computation for the MC interpretation and the bottom right the network for the RR interpretation. The networks are shown for the input The witness examined .... The arg2 node is yet to be instantiated and so has no values assigned. The two competing interpretations condition the lexical network (shown unconditioned on the top figure) with different constraints. For instance MC interpretation requires that the Subject NP be the Agent. These conditioned values are shown in boldface font. The overall posterior is computed based on the network parameters and the conditioning values.

$P(I)^t$	P(Frame|V)	P(Fit|V, Subj)
$P(I = MC)$	P(transitive|verb=arrest)	P(Fit=Agent|subject=crook,verb=arrested)
$P(I = RR)$	P(intranstive|verb=arrest)	P(Fit=Theme|subject=crook,verb=arrested)
$P(I = MC)$	P(transitive|verb=examine)	P(Fit=Agent|subject=witness,verb=examine)
$P(I = RR)$	P(intransitive|verb=examine)	P(Fit=Theme|subject=witness,verb=examine)
$P(I = MC)$	P(transitive|verb=race)	P(Fit=Agent|subject=horse,verb=raced)
$P(I = RR)$	P(intransitive|verb=race)	P(Fit= Theme|subject=horse,verb=raced)

Table 1: Constraints on parameters for the lexical valence probability computation for different interpretations (MC and RR). The table above shows the lexical valence structures for the sentences The crook arrested ..., The witness examined ..., and The horse raced ...

instance, MC requires the subject NP to be an $Agent$, while RR requires that the $Frame$ variable be set to the value $transitive$ and the subject NP to be a $Theme$. The MC interpretation could either have a transitive or intransitive frame, so there is no constraint imposed on the subcategorization frame for this interpretation.

How do these constraints play a role in the evaluation of the networks to compute the posterior support for the two interpretations (MC and RR)? To illustrate this, we now go through a simplified evaluation of the two networks to compute the MC and RR posteriors after the input "The witness examined" (see Figure 13, bottom row). For the MC interpretation, the thematic posterior, MC^t_{thm} (t is the index into the specific stage where the posterior is computed (after "NP V")) is[6]

$$P(V, Arg1 = Subj, Frame, Fit(Arg1, V) = Agent|V = examine, Subj = witness) =$$
$$P(Fit = Agent|V = examine, Subj = witness) \sum_{Frame} (Frame|V = examine) =$$
$$P(Fit = Agent|V = examine, Subj = witness) \quad (43)$$
$$(44)$$

[6]For ease of exposition, we don't consider the yet unseen arg, $Arg2$. It's effect on the posterior at this stage is the same for the two interpretations.

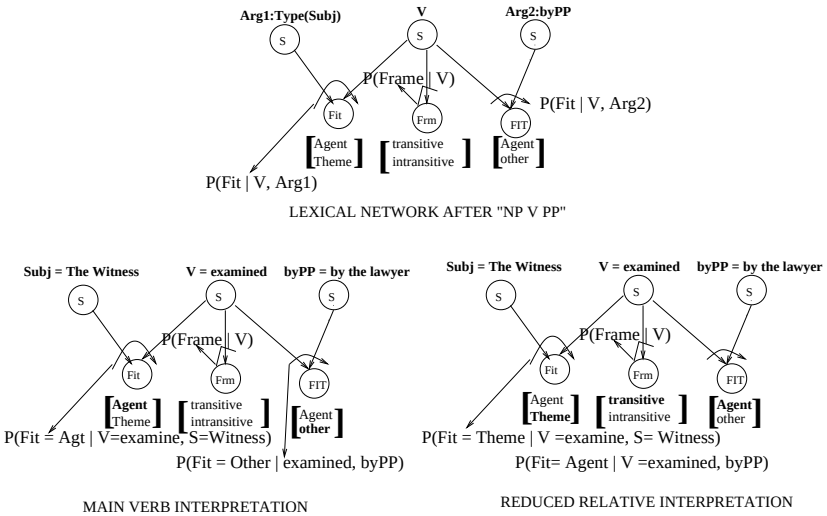

Figure 14: Bayes Nets for valence Probabilities for the McRae example at a later stage (compared to Figure 13 after the "byPP" input has been processed. Here $Arg2$ is the byPP for the input "by the lawyer" phrase. The bottom left shows the network computation for the MC interpretation and the bottom right the network for the RR interpretation. The networks are shown for the input The witness examined by the lawyer.... The by phrase and the second NP (the lawyer) are additional input nodes that influence the thematic fit and the posteriors. For instance RR requires that the second NP be the Agent.

The MC interpretation does not constrain the $frame$, since it can apply to both transitive and intransitive verbs. To take account for this fact, we sum over all it's values. This summation (marginalizing over the $frame$ variable) for a given verb sums up to 1 and hence gets taken outof the final equation.

The RR interpretation, however, requires the verb to be transitive **and** the subject to be the theme. Hence, here we have the two constraints (shown in boldface in Figure 13. With these, constraints, the reduced relative interpretation thematic posterior, RR^t_{thm} is

$$P(V, Arg1 = Subj, Frame = transitive, Fit(Arg1, V) = Theme | V = examine, Subj = Witness) = P(Fit = Theme | V = examine, Subj = witness) \times P(Frame = transitive | V = examine) \quad (45)$$

Our second study modeled the sentential compliment SC versus direct object DO behavioral data reported in Pickering *et al.* (2000) (recall "The athelete realized her ..." examples from the previous section and from the introduction). Table 2 shows the parameters for the different interpretations (Sentential Complement (SC) and Direct Object (DO)) after the input "the young athlete realized her potential". Details of the network structure and model can be found in Section 4.3.

$P(I)$	P(Frame|V)	P(Fit|V, Subj)
$P(I = DO)^t$	P(DO(frame)|verb=realize)	P(Fit=Agent|subject=athlete,verb=realized)
$P(I = SC)^t$	P(SC(frame)|verb=realize)	P(Fit=Theme|subject=athlete,verb=realized)
$P(I = DO)^{t+1}$	P(DO(frame)|verb=realize)	P(Fit=Theme|subject=athlete,verb=realized,NP= her potential)
$P(I = SC)^{t+1}$	P(SC(frame)|verb=realize)	P(Fit=Proposition|subject=athlete,verb=realized,NP= her potential)

Table 2: Constraints on parameters for the lexical valence probability computation for different interpretations (DO and SC). The table above shows the lexical valence structures for the sentences The young athlete realized her potential... at two stages; one before and one after the NP " her potential".

2.2.7 Other estimators

Figure 5 shows our general architecture involving different evidential sources (top-down and bottom-up) that contribute different degrees of support for an interpretation. We described in detail three of the simplest estimators which as we show in the following sections suffice to model important aspects of human sentence processing. Of course, as we are able to investigate more subtle aspects of sentence and discourse processing, we fully expect (indeed as Figure 5 suggests), other estimators including discourse and deeper semantic sources to become increasingly necessary and important. We further believe that these sources exhibit significant structure and the techniques for building structured estimators (as for the syntactic and lexical/thematic sources) and combining them (described in the next section) provide a flexible and natural framework to investigate their contributions to language interpretation. We (and we hope others) will use our framework for adding new knowledge sources, making predictions about reading times and other even finer-grained aspects and testing their validity experimentally.

2.3 Combining probability estimators

The last section outlined how we calculate the various probabilistic components (the syntactic, lexical valence, and word N-gram) of our Bayesian model. Of course, all of these (and possibly other) components have to be combined to provide an estimate of the total posterior probability for a given interpretation.

In some cases, as with the SCFG, we have relatively complete models of the independence assumptions between probabilities. In other cases, for example between thematic and syntactic probabilities, we do not yet have a good idea what the exact causal relationship is between probabilities.

Ideally, we would like the combination technique to be independent of the data domain, so we can avoid creating one rule for the interaction of syntax with semantics and another for the interaction of syntax with lexical valence and yet another for the interaction of syntax with word N-grams etc.

Fortunately, there is a canonical and widely applicable model of probabilistic source combinations that works for our purpose. The model is called a NOISY-AND model (Pearl, 1988) which is the method of choice when a member of a set of several components (say the syntactic component) can cause a specific outcome (in this case a specific interpretation to be selected), and where the likelihood of the outcome is very high only when all the conditions prevail simultaneously. The NOISY-AND model (Pearl, 1988) is thus a causal independence assumption made in computing the *conjunctive impact* of the multiple sources. Furthermore there is now good evidence from development studies that this model seems to be an important inductive bias in causal learning in children (Cheng 1997).

The NOISY-AND model makes the following two assumptions.

1. **Accountability**: An Event E is false if any of the causal factors is false.

2. **Enabling Independence**: If both conditions C1 and C2 can cause an Event E, then the mechanism that disables the effect of C1 on E is independent of the mechanism that disables the effect of C2 on E.

The NOISY-AND model is the probabilistic interpretation of the logical AND. In the model, each parent X_i, a binary stochastic variable, is interpreted as the condition for the effect Y (also a binary stochastic variable). So in the case in Figure 15, the various parents are the different types of support for the competing interpretations I_1 and I_2.[7]

The NOISY-AND requires that the enabling effects to be independent. Let us assume that the effect of an individual source, $S_i \in S$, is characterized by the enabling probability $p_{s_i} = P(I = t|S_i = t)$. Then, using the NOISY-AND causal independence assumption, the interpretation is true to the extent that the enabling sources (s) are active.

$$P(I = T) = \prod_{S_i = t} p_{s_i} \tag{46}$$

$$P(I = F) = 1 - \prod_{S_i = t} p_{s_i}$$

[7] There is also a generalized version of NOISY-AND called NOISY-MIN that allows for multiple (non-binary) interpretations to be simultaneously considered.

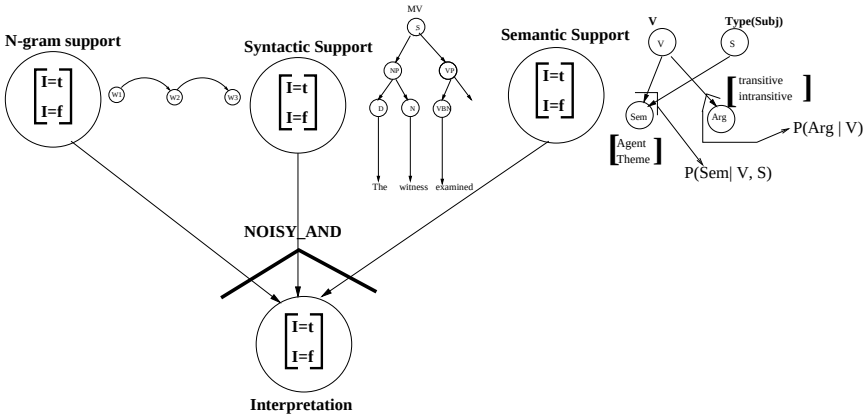

Figure 15: The NOISY-AND Model to combine multiple conjunctive sources for an interpretation I.

So for a set of m sources s_1 to s_m(with a posterior probabilities $P(S_1 = t) \ldots P(S_m = t)$, we have the following equation.[8]

$$P(I = T) = \prod_{s=1}^{s=m} P(S_i = t) \times p_{s_i} \tag{47}$$

Notice that is we set uniform weights of 1 for the influence of the individual sources ($p_{s_i} = 1; \forall s_i \in S$), the equation above becomes a multiplication of the posterior probabilities of the individual sources.

$$P(I = T) = \prod_{s=1}^{s=m} P(S_i = t) \tag{48}$$

For all the experiments described in this paper, we set $p_{s_i} = 1$.

In our model, we need to compute the posterior probability of each interpretation $I_i \in (I_1 \ldots I_n)$. Since human parsing is incremental, we will need to re-compute this probability after each input stage, i.e. after each word $t_i \in (t_1 \ldots t_k)$. Thus the preferred interpretation at each stage is the interpretation which maximizes this posterior probability. In other words, the preferred interpretation at time t, $I*^t$, is:

$$P^*(I^t) = \underset{i \in \text{interpretations}}{\text{argmax}} \quad P(I_i^t) \tag{49}$$

How are each of these posterior probabilities of parses computed? The posterior probability is a NOISY-AND of all the different types of support for the interpretation (including lexical, syntactic and valence support). Let s range over the m various types of support for an interpretation; $s \in (syntax, lexical, valence), m = 3$. Then $P(I_i^t)$, the probability of an interpretation i, can be computed as follows:

$$P(I_i^t) = \frac{\prod_{s=1}^{s=m} I_{i_s}^t}{\sum_{j=1}^{j=n} \prod_{s=1}^{s=m} I_{j_s}^t} \tag{50}$$

In other words, for each type of evidential support for an interpretation we separately compute the probability of the interpretation given that support, and then sum and normalize.

[8]Recall that in our sentence processing model the posterior probabilities of the various sources are the output probabilities of the estimators described in the previous section.

How are the probabilities given each of these types of evidence computed? We've discussed 3 types of evidence: lexical N-gram, syntactic, and valence. The posterior probability given each kind of evidence can be expressed as follows:

$$P(I^t_{i_{\text{lex}}}) = P(I^t_i|(\text{ngram-model})w_1, w_2, ..., w_t) \tag{51}$$

$$P(I^t_{i_{\text{syn}}}) = P(I^t_i|(\text{syntactic-model})w_1, w_2, ..., w_t) \tag{52}$$

$$P(I^t_{i_{\text{val}}}) = P(I^t_i|(\text{valence-model})w_1, w_2, ..., w_t) \tag{53}$$

$$\tag{54}$$

To summarize, the best interpretation at time t, $P^*(I^t)$, can be computed as:

$$P^*(I^t) = \underset{i \in \text{interpretations}}{\text{argmax}} \frac{P(I^t_{i_{\text{lex}}}) \times P(I^t_{i_{\text{syn}}}) \times P(I^t_{i_{\text{val}}})}{\sum_{j=1}^{j=n} \prod_{s=1}^{s=m} I^t_{j_s}} \tag{55}$$

Thus the preferred interpretation is the one which has the maximum posterior probability given all the evidence.

3 The Predictions of our Model

The previous section described how the model assigns probabilities to different parses of sentences. In this section we describe how the probabilities, and the on-line updating of these probabilities as each new word is read, affect behavioral performance.

The first kind of behavior that the model predicts is parse preference. As the previous section described, the preferred interpretation is the one which has the maximum posterior probability given all the evidence. The previous section also described how this probability is computed. Thus the prediction of the model is that the preferred interpretation at any point in the processing of a sentence is the interpretation with the highest posterior probability.

The remainder of this section focuses on a further behavioral prediction: processing time. We will describe two predictions about how long it takes to read words or phrases in the context of particular ambiguities.

3.1 Reading time: the role of the Expectation principle

Our first reading time prediction has already been sketched in Equation 22 in the introduction, and is based on the expectation principle. This principle states that the parser implicitly maintains probabilistic expectations about upcoming words and structure, and that the parser assumes that future words will be consistent with these probabilities. Words which violate these expectations produce increased reading time.

Equation 22, repeated below as equation 56, gives the heart of the proposal.

$$\text{reading time}(word) \propto \frac{1}{P(\text{word}|\text{context})} \tag{56}$$

In order to operationalize this proposal, we have to flesh out the term $P(\text{word}|\text{context})$. The conditional probability of a word given the previous context can be expressed as follows:

$$P(w_i|w_1, w_2 \cdots w_{i-1}, \text{parsetree}(w_1, w_2 \cdots w_{i-1}), \text{valence}(w_1, w_2 \cdots w_{i-1})) \tag{57}$$

Following (Hale, 2001), we can use the definition of conditional probability to re-write this equation for conditional probability as the ratio of two joint probabilities. This rewrite makes it clear that the conditional probability of a word given the context is related to the *change* in probability caused by the introduction of a new word. As in the previous section, we use $P(I^t)$ to mean the probability of interpretation I at time t:

$$
\begin{aligned}
P(w_i|w_1, w_2 \cdots w_{i-1}, & \\
\text{parsetree}(w_1, w_2 \cdots w_{i-1}), & \\
\text{valence}(w_1, w_2 \cdots w_{i-1})) & \\
&= \frac{P(w_1 \ldots w_i)}{P(w_1 \ldots w_{i-1})} \\
&= \frac{P(I^t)}{P(I^{t-1})} \qquad (58)
\end{aligned}
$$

Thus the conditional probability of a word can instead be expressed as a ratio of two probabilities: the probability of interpretation I at time t divided by the probability of interpretation I at time $t-1$. But this equation relies on the simplifying assumption that each sentence only has one interpretation. Of course this isn't true, and since our model has a parallel architecture, more than one interpretation may be maintained at any time. Recall from the previous section that we therefore need to compute the posterior probability of each interpretation $I_i \in (I_1 \ldots I_n)$ for each input stage $t_i \in (t_1 \ldots t_k)$, and then combine these via NOISY-AND:

$$P(I_i^t) = \frac{\prod_{s=1}^{s=m} I_{i_s}^t}{\sum_{j=1}^{j=n} \prod_{s=1}^{s=m} I_{j_s}^t} \qquad (59)$$

where s ranges over the m various types of support for an interpretation; $s \in (syntax, semantics, lexical, thematic)$.

How does the ration of probabilities we have discussed lead to a claim about processing time? Let us define a variable corresponding to the change in probability caused by the introduction of a new word, called δ, and allow in its definition the possibility of multiple interpretations:

$$\delta(I_{i^t}) = \frac{P(I_i^t)}{P(I_i^{t-1})} \qquad (60)$$

We can now give a flesh out the intuition of Equation 56 as follows:

$$\text{ProcessingTime}_{\text{Expectation}} \approx -\delta(I_{i^t}) \qquad (61)$$

3.2 Reading Time: the role of the Attention Principle

The second reading time prediction comes from the Attention principle, which states that although the comprehension mechanism may keep multiple parallel interpretations, that the attentional focus is on the most-probable interpretation. Any time this highest-ranked interpretation drops from its high position, the surprise causes a longer reading time. One way that a re-ranking can cause a processing delay is when a new word is read which lowers the probability of the first interpretation more than the second-ranked interpretation, causing the two to become reordered, or 'flipped' in preference. Since the first interpretation has attentional focus, attention must shift whenever some other interpretation replaces this one, causing a processing delay.

The mathematical definition of reordering is quite simple. Recall that $P^*(I^t)$, the most preferred interpretation at time t, is defined as the interpretation which has the maximum posterior probability:

$$P^*(I^t) = \underset{i \in \text{interpretations}}{\text{argmax}} \; P(I_i^t) \qquad (62)$$

A reordering is then defined as a change in preferred interpretation:

$$P^*(I^t) \neq P^*(I^{t-1}) \qquad (63)$$

Any reordering of this kind causes a reading time increase. What is the magnitude of this increase? The expectation principle predicts a reading time increase proportional to the change in probability mass, $\delta(I_{i^t})$. What additional

increase in processing time should be accounted for by the reordering? We can't know the amount of this increase in advance, but we make the simplifying assumption that it is a linear function of the expectation-based reading time. Let w_{flip} be a weight which indicates the additional increase do to reordering. The following equation then gives our prediction for the impact on processing time.

$$\text{ProcessingTime}_{\text{Reordering}} \approx$$
$$= \begin{cases} -w_{flip} \times \delta(I_{i^t}), & \text{if } P^*(I^t) \neq P^*(I^{t-1}) \\ -\delta(I_{i^t}), & \text{if } P^*(I^t) = P^*(I^{t-1}) \end{cases}$$

While we are not able in this paper to exactly determine the proper value for w_{flip}, we will show later that simply setting w_{flip} to 2 accurately accounts for behavioral results.

4 Motivating Examples: preference handled by a probabilistic model

Before we turn in the next two sections to test our probabilistic model against behavioral results from (McRae *et al.*, 1998) and Pickering *et al.* (2000), we use this section to show how the probabilistic model can be used to explain parse preferences due to probabilistic structure. We choose two simple example; preference due to morphological category probability, and preference due to syntactic category probability.

4.1 Morphological Category Probability

We know that the frequency of words and in particular of the different morphological or syntactic categories of a word plays a role in parse preference. For example Burgess and Hollbach (1988) and Trueswell (1996) studied words such as *searched* and *selected*, which are ambiguous between a preterite (simple past) and a participle reading (sometimes called the VBD/VBN ambiguity after the respective names for the preterite and participle part-of-speech tags in the Penn Treebank tagset). Verbs like *selected* are more likely to be a participle, while *searched* is more likely to be a simple past, as shown in the following table:

Selected: 89% participle, 11% simple past
Searched: 22% participle, 78% simple past

Trueswell (1996) showed that these more fine-grained lexical category probabilities play a role in the disambiguation of main verb/reduced relative ambiguities. He did this by embedding these verbs in sentences which have a local ambiguity. Each sentence had an initial word sequence like *The room searched* which is syntactically ambiguous between a relative clause reading (compatible with the participle form) and a main-verb reading (compatible with the simple past). Trueswell found that verbs with a frequency-based preference for the simple past form caused readers to prefer the main clause interpretation (as measured by longer reading time for a sentence which required the other interpretation such as (64)):

(64) The room searched by the police contained the missing weapon.

This suggests that the frequency with which the different morphological categories of a verb occur plays a role in whether one syntactic parse is preferred or not.

How does the Bayesian model handle the results of Trueswell (1996), that showed an effect of lexical category frequency on preference? The fact that, for example, the word *selected* is more likely to be a participle than a simple past, while *searched* has the opposite preference, is handled in our model by the probabilistic syntactic grammar. In the SCFG, this is represented by the fact that *P(VBN|selected)* is higher than *P(VBD|selected)*. The SCFG tree structure includes the likelihood P(selected|VBN); we can use Bayes rule on the SCFG structure to compute the proper posterior (counts are again from the Brown corpus):

$$P(VBN|selected) = \frac{P(\text{selected}|VBN)P(VBN)}{P(\text{selected})} = .0022 * .029/.000071 = .90 \quad (65)$$

$$P(VBD|selected) = \frac{P(\text{selected}|VBD)P(VBD)}{P(\text{selected})} = .00015 * .047/.000071 = .10 \quad (66)$$

(67)

The SCFG structure thus lets us compute that the posterior probability of *selected* being a participle (VBN) is thus 9 times higher than its probability of being a preterite (VBD).

4.2 Syntactic Structure Probability

There is also evidence that the probability of larger (supralexical) syntactic structures plays a role in processing. We saw earlier that McRae *et al.* (1998) used the low probability of the reduced relative structure, as compared with the main clause structure, as part of their model of reduced relative clause difficulty. In this section we briefly explore another kind of ambiguity: sentences beginning with an embedded sentential subject, which are known to cause processing problems. For example, the word *that* is ambiguous between a determiner and a (more frequent) complementizer. Consistent with work described above on lexical category frequencies, Juliano and Tanenhaus (1993) found that *that* is interpreted most easily as a complementizer after verbs. But sentence-initially, when interpreting the the word *that* as a complement would require an embedded sentential subject, *that* is instead interpreted as a determiner.

In the following sentences from Juliano and Tanenhaus (1993), for example, readers incorrectly parse the word *that* as a complementizer in 68, causing an increase in reading time at the word **diplomat**. Similarly, readers incorrectly parse the word *that* as a determiner in 68, causing an increase in reading time at the word **diplomats** in (71). Sentences that are compatible with readers preferences are underlined.

(68) The lawyer insisted *that* experienced **diplomat would** be very helpful

(69) The lawyer insisted *that* experienced **diplomats would** be very helpful

(70) *That* experienced **diplomat would** be very helpful to the lawyer.

(71) *That* experienced **diplomats would** be very helpful made the lawyer confident.

The SFCG model successfully predicts Juliano and Tanenhaus's (1993) result that the word *that* tends to be interpreted as a complementizer after a verb, but as a determiner at the beginning of a sentence. The dispreference for an initial complementizer, for example, follows from the very low probability of rules like S→SBAR VP. This rule, corresponding to a sentential subject of the main clause, has an extremely low probability (.00065). Partial SCFG parses for the four sentences are shown in Figures 4.2– 19.

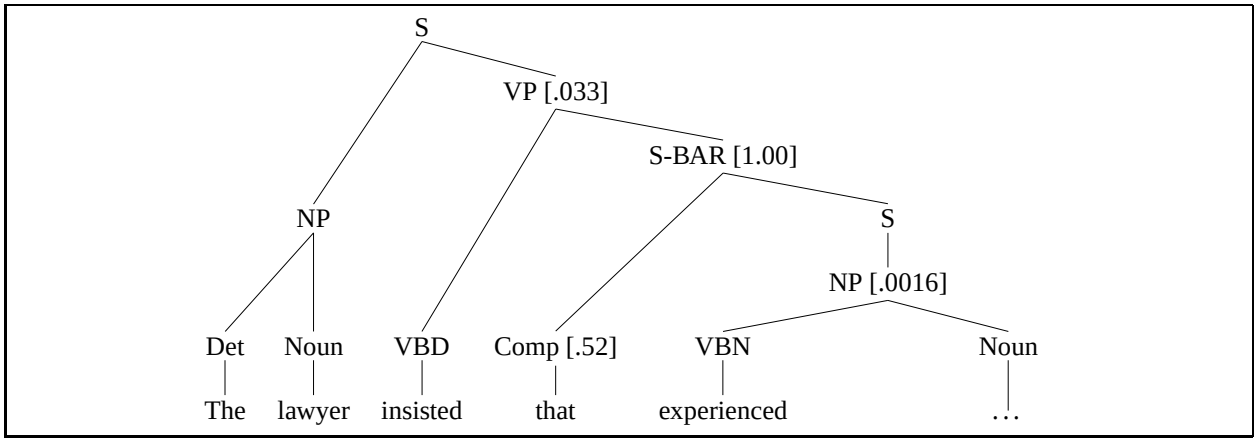

Figure 16:

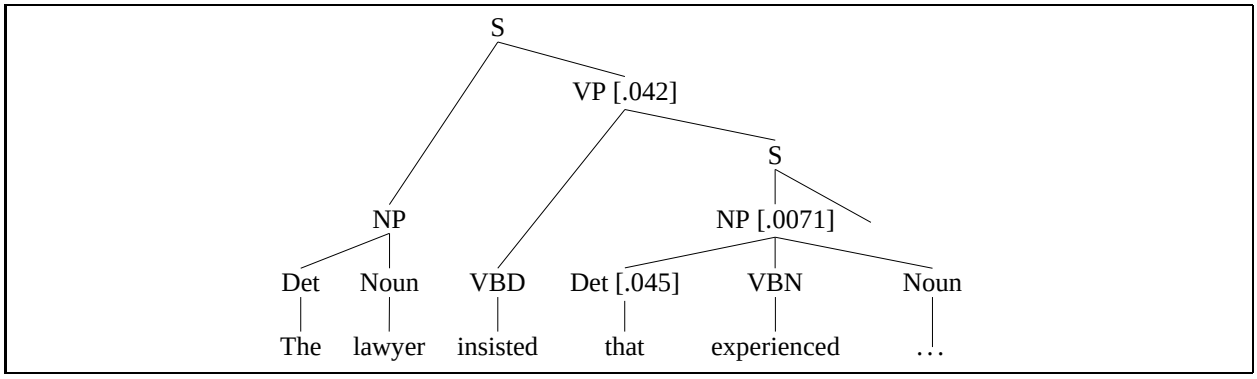

Figure 17:

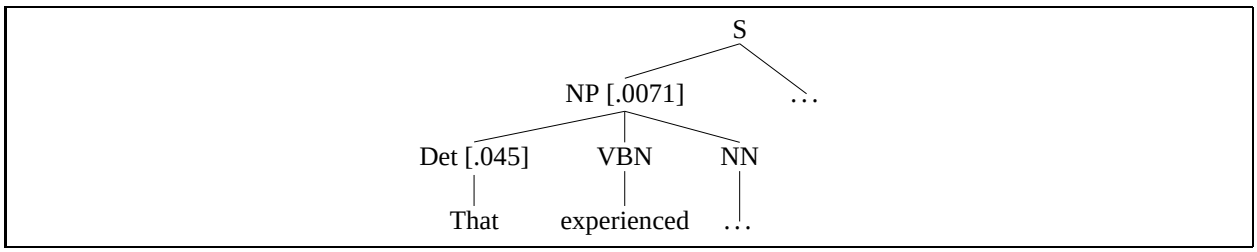

Figure 18:

In each of these two pairs of parses, we have shown the rule probabilities only where the rules are different across parses. For example, and show the initial parse fragments for the two possible parses of the sentence fragment *The lawyer insisted that experienced…*. At this point, the parse trees only differ in 7 rules. We have augmented the tree with the probabilities for each of these rules. Our model's preference for the parse in which *that* is a COMP can be computed by multiplying the probabilities of the various rules that are unique to each parse.

```
P(parse in which that = COMP) = .033 * .52 * .0016 = .000027456
P(parse in which that = DT) = .042 * .0071 * .045 = .0000134190
```

Similarly, Figure 18 and Figure 19 show the initial parse fragments for the two possible parses of the sentence fragment *That experienced*. At this point, the parse trees only differ in 6 rules. (the second parse, in which *that* is a COMP, is more complex, with 2 more rules than the first parse). Once again, we have augmented the tree with the probabilities for each of these rules. In this case, our model assigns a higher probability to the parse in which *that* is a DT (Determiner). This probability can be computed by multiplying the probabilities of the various rules that are unique to each parse.

5 Study One: The Main Clause/Reduced Relative Ambiguity and McRae *et al.* (1998)

The results in the previous section sketches the intuitions of how our probabilistic model is accounts for qualitative results on disambiguation preference. In this section we test the model more carefully by seeing if can account for the results of a comprehensive reading time experiment. As described earlier, we chose to model the data collected by McRae *et al*. (1998) for two reasons. First, it is crucial to show that our model can handle a wide variety of well-studied cases of ambiguity, and the MC/RR is perhaps the most-studied case. Second, the (McRae *et al.*, 1998) study provides a model for their results based on their own competition-based model. Since they provide the norms and counts that their model is trained on, this allows us to compare more directly against their model.

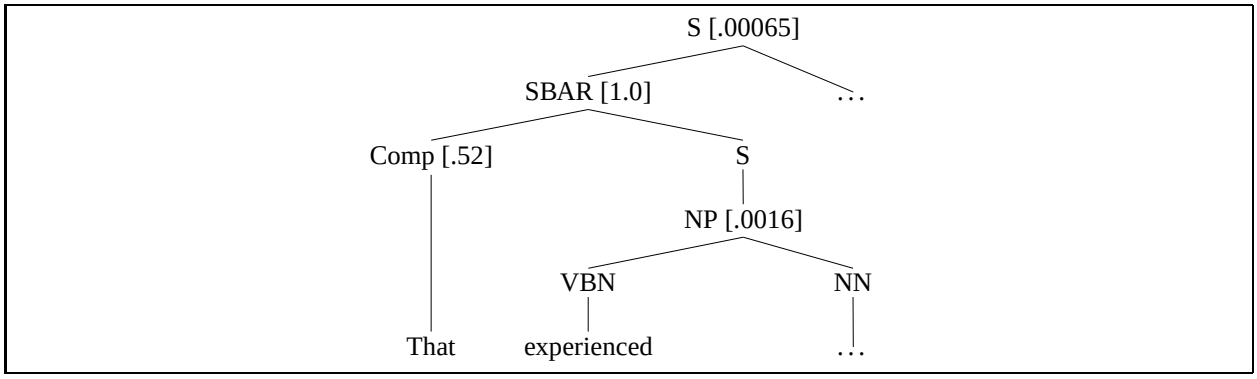

Figure 19:

5.1 The Model and the Input Probabilities

The model is based on the Bayes net described in Section 4. The parameters thus consist of the structure of the net expressing probabilistic independence assumptions, and the probability tables associated with the net. The overall structure of the combined lexical-thematic and syntactic networks for the MC vs. RR interpretations is shown in Figure 20.

As shown in Figure 20, there are two Bayes nets computing the semantic and syntactic fit of the input sentence to the different possible interpretations.

1. The column on the left in Figure 20 computes the support provided by the thematic role and semantic fit for the two interpretations (the Main Clause (MC) (top row) and Reduced Relative (RR)(bottom row)). The structure and parameters of this (sem) network are similar to the one shown in Figure 8.

2. The column on the right in Figure 20 computes the support provided by the syntactic parses of the input sentence for the two interpretations (the Main Clause (MC) (top row) and Reduced Relative (RR) (bottom row)). The structure and parameters of this (syn) network are similar to the one shown in Figure 7.

The combined evidence for a particular interpretation is obtained by taking the NOISY-AND that estimates the conjunctive support of the different sources.

The various parameters of the network in Figure 20 are the conditional probabilities both for the syntactic and the valence/subcategorization networks. The syntactic probabilities are computed based on an SCFG grammar and the network structure and computations for the SCFG parse are explained in Section 2. The probabilities are described in Table **??**. For the valence network, we used the probability of the initial NP being an Agent (Patient) given the verb and initial NP P(Agent (Patient)|verb, initial NP). These numbers were obtained from the norming studies reported in McRae *et al.* (1998).

The first row in Table 3 expresses the probabilistic constraint that the word “cop”(for example) is an agent, given that the verb is “arrested”. The second row constraint expresses the probability that it is a patient. For both these, we used the norming study reported in McRae *et al.* (1998), where subjects rated the plausibility of the word “cop” as an agent and as a patient of the predicate arrest on a scale of 1 to 7. We used the norming scale as a conditional probability. So, if the agent was rated 4 on the scale, we took the P(Agent|verb, initial NP) in that case to be $4/7$. In general, when we had norming study data, we used this approximation to conditional probabilities.

The third and fourth constraints express the probability that the ”-ed” form of the verb is a participle versus a simple past form (for example P(Participle|“arrest”)=.81). These were computed from the POS-tagged British National Corpus. Verb transitivity probabilities were computed by hand-labeling subcategorization of 100 examples of each verb in the TASA corpus. (for example P(transitive|“entertain”)=.86). Main clause prior probabilities were computed by using an SCFG with rule probabilities trained on the Penn Treebank version of the Brown corpus. Section 3 and Section 5 detail the SCFG probability calculation procedure.

Appendix C summarizes the Good Agent, Good Patient probabilities (from the norming study), the transitive versus intransitive bias at the specific verb and the Simple Past (needed for the Main Verb interpretation) versus Past Participle distinction calculated for the 40 verbs in the McRae study.

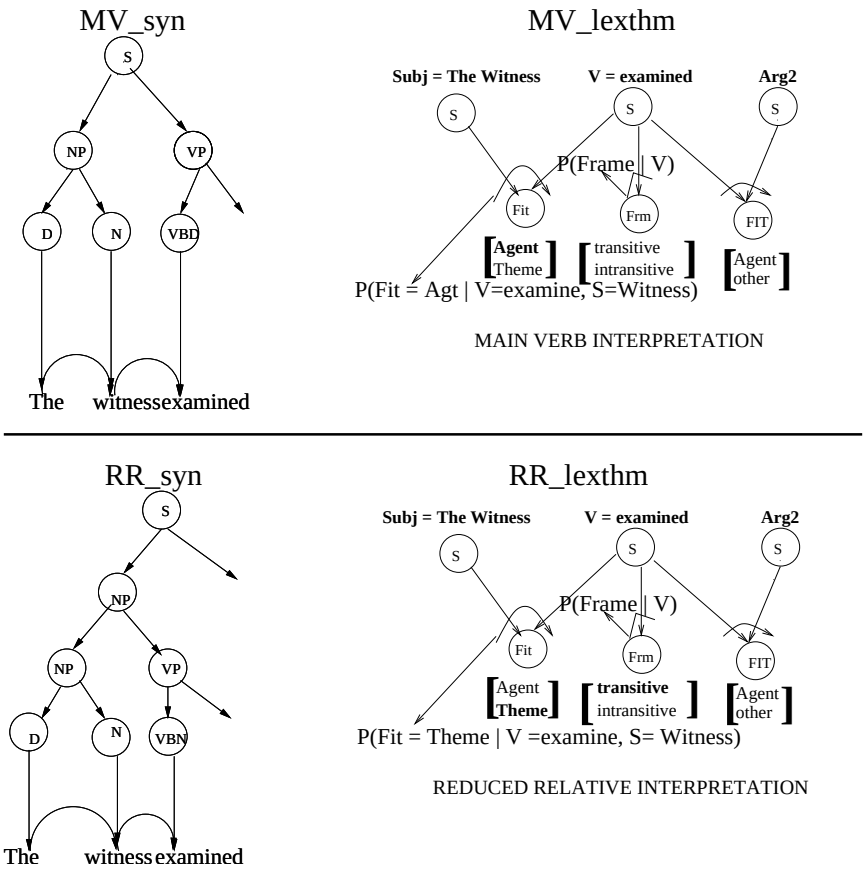

Figure 20: A Bayes net combining SCFG probabilities (*syn)* with subcategorization, thematic (*thm*), and other lexical probabilities to represent support for the main verb (MC) and reduced relative (RR) interpretations of a sample input.

From the parameters, for each sentence in the McRae data, we compute the lexical, semantic and syntactic support for the two interpretations MC and RR. For each stage in the input

At the initial NP (ex. the witness),
At the verb (ex. examined),
At the preposition (ex. by),
At the second NP (ex. the lawyer),

our model computes the following entities

$$P(MC^t|syn), P(MC^t|sem), P(RR^t|syn), P(RR^t|sem) \tag{72}$$

Our model computed the SCFG based syntactic probability and the thematic and semantic fit probabilities for the MC and RR interpretations at different points in the input. These probabilities were then combined using the NOISY-AND function described in the previous section. So, for each input stage, we computed the posterior value for the MC and RR interpretation given the syntactic and the thematic/semantic support.

5.2 Model results

We tested our model on sentences with the 40 different verbs in McRae *et al.* (1998). For each verb, we ran our model on sentences with Good Agents (GA) and Good Patients (GP) for the initial NP. Our model results are consistent

Data	Source
Valence Probabilities	
Valence Probabilities for the Subject NP	
P(Agent|verb, initial NP)	McRae *et al.* (1998)
P(Patient|verb, initial NP)	McRae *et al.* (1998)
P(transitive|verb)	TASA corpus counts
P(intransitive|verb)	TASA corpus counts
Valence Probabilities for the PP Agent	
P(RR|initial NP, verb-ed, by)	McRae *et al.* (1998) (.8, .2)
P(RR|initial NP, verb-ed, by,the)	McRae *et al.* (1998) (.875. .125)
P(Agent|initial NP, verb-ed, by, the, NP)	McRae *et al.* (1998) (4.6 average)
SCFG Probabilities	
P(MC| SCFG prefix)	SCFG counts from Penn Treebank and BNC
P(RR| SCFG prefix)	SCFG counts from Treebank and BNC
P(Participle|verb)	SCFG counts from Treebank and BNC
P(SimplePast|verb)	SCFG counts from Treebank and BNC

Table 3: Source of probabilities for our model

with the on-line disambiguation studies with human subjects (human performance data from McRae *et al.* (1998)) and show that a Bayesian implementation of probabilistic evidence combination accounts for garden-path disambiguation effects. We first walk through how the model assigns probabilities to two sentences. We then test the Bayesian model against the behavioral results on sentence completion from McRae *et al.* (1998). Finally, we test our model against the behavioral results on reading time from McRae *et al.* (1998).

5.2.1 Walking through the assignment of probabilities to two sentences

Table 5.2.1 refers to the assignment of probabilities by our model to the two sentences:

(73) The *witness*/ examined by / the lawyer / turned out / to be unreliable.

(74) The *evidence*/ examined by / the lawyer / turned out / to be unreliable.

Examine	Init NP	verb-ed	by	the	agent NP
$P((MC)/P(RR)\|GA)$	2.91	1.729	.432	.062	.01
$P((MC)/P(RR)\|GP)$	0.47	0.201	.090	.039	.01

Table 4: $P(M)/P(R)$ results of the model on example sentences "The *witness* examined by the lawyer turned out to be unreliable (Good Agent (GA)), and "The *evidence* examined by the lawyer turned out to be unreliable" (Good Patient (GP)).

Shown in Table 5.2.1 is the ratio of the posteriors ($P(MC/P(RR))$ for the Main Clause (MC) and Reduced Relative (RR) interpretations for the two sentences. The difference in the sentences is that in one case, the Subject NP, *witness*, is animate (and hence a Good Agent (GA); while in the other case the subject NP, *evidence*, in inanimate and hence a Good Patient (GP). The ratio of the posteriors is computed at various points in the input, such as at the initial NP, the verb, after the preposition "by" and the determiner "the", and at the agent N "lawyer".

- At the end of the initial NP, the MC interpretation is more 2.9 times more likely as the RR interpretation for the GA (the witness) and .47 times as likely for the GP (*evidence* case.
- At the main verb boundary (*examined*) the ratio of the posteriors changes to be ($P(MC)/P(RR) = 1.73$) for the GA sentence and ($P(MC)/P(RR) = .20$) for the GP sentence.

- At the next stage, which is the "by" phrase, our model continues the expected trajectory for the GP sentence and the MC interpretation continues to decrease from being 20% as likely as the RR interpretation to being 10% as likely. For the GA sentence, however, we see an expectation violation, in both a sharp decline for the MC interpretation as well as a demotion of the best interpretation; a *flip* in preference for the two interpretations.

- At the verb boundary the MC interpretation was the prefered one $(1.73 : 1)$. After the "by" phrase is encountered, the prefered interpretation changes to RR and the MC interpretation is now only half as likely as the RR interpretation. Thus based on the posterior probabilities, there is an violation of expectation. The best interpretation has become disprefered and the second-best interpretation is now the leading candidate interpretation. Thus our model predicts reading time difficulty for the GA sentence at the "by" phrase and none for the GP sentence.

5.2.2 Modeling the Sentence Completion Study of McRae *et al.* (1998)

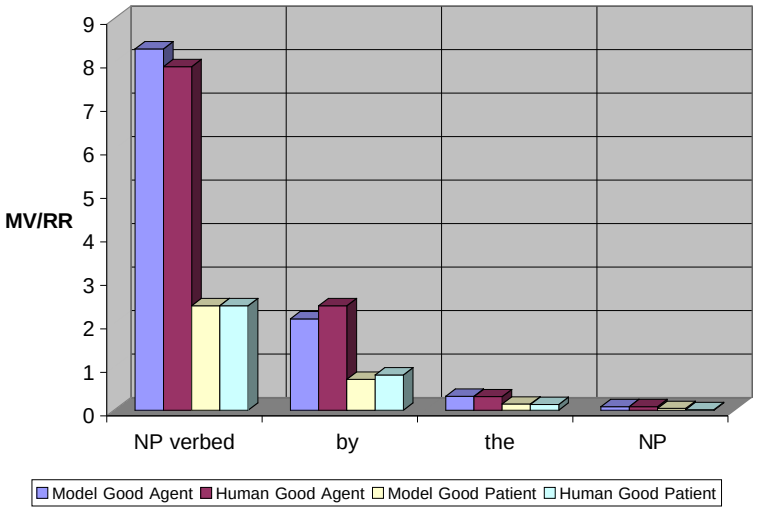

Figure 21: Human sentence completion results (MC counts/RR counts) and model predictions (P(MC)/P(RR) for the McRae *et al.* (1998) sentence completion data.

Figure 21 shows the predictions of our model as well as the human sentence completion data from the McRae *et al.* (1998) experiment. The human and model predictions were computed at four stages:

1. the *verb* (*The crook arrested*),
2. *by* (*The crook arrested by*),
3. *the* (*The crook arrested by the*)
4. the Agent NP (*the crook arrested by the detective*).

For the human data, the Y axis in Figure 21 shows the ratio of sentence completion count (MC counts/RR counts). For the model, the Y axis in Figure 21 shows the ratio of probabilities P(MC)/P(RR).

The human data (the second and fourth bars at each word in Figure 21) shows a number of trends. First, thematic fit clearly influenced this gated sentence completion task. Note that the Good Agent sentences have a higher MC/RR ratio at the "NP verbed" stage than the Good Patient sentences . The model matches this difference. Next, at the "by phrase", the human data shows that the posterior probability of producing an RR interpretation increases sharply (hence the MC/RR ratio drops). Thematic fit is at least one of the factors influencing this increase, since the Good Agent MC/RR ratio is still higher than the Good Patient MC/RR ratio. Finally, both the model and the human data reliably predict that after seeing the second NP, there is no chance of generating an MC completion, since the MC/RR ratio has gone to zero.

5.2.3 *Flips*: A Qualitative correlate of reading time effects

Our model predicts a qualitative difference in reading times when there is a reranking through demotion of the top (highest posterior) interpretation to a lower rank. We refer to this situation as a *flip*. Figure 22 shows the change in the posterior probability for the MC and RR interpretations for the Good Agent (GA) cases. The data is averaged over the 40 verb/GA sentences in the McRae data. The data shows the following effects

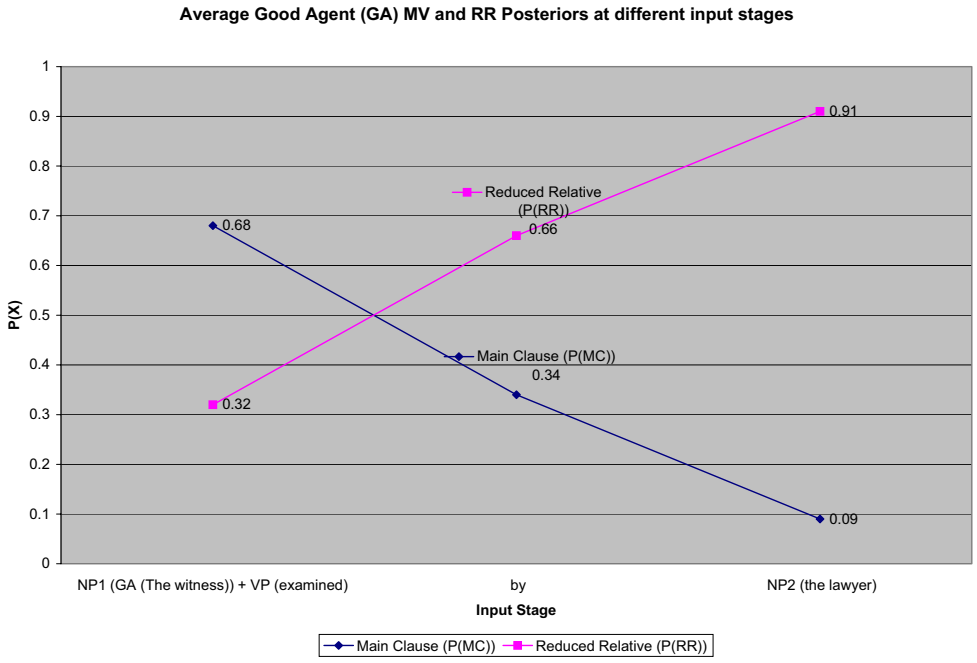

Figure 22: P(MC) and P(RR) for the ambiguous region showing a flip

1. At the initial NP phrase the MC interpretation is twice as likely as the RR interpretation. This reflects the $SCFG$ structural probabilities prior to seeing the verb (not shown in figure).

2. At the verb boundary phrase the MC interpretation is still high (more than twice the RR interpretation). This reflects the fact that although the verbs are likely to reflect a high transitive bias (favoring the RR interpretation), the fact that the subject is a good agent continues to favor the MC interpretation. The combined effect is reflected in the average posterior probability of the MC interpretation which is now only twice as probable as the RR probability.

3. After the "by" phrase, things change a lot. Now we notice that the RR posterior is twice as high as the MC posterior. This reflects the RR bias at the by phrase, where there is now a high probability that the initial NP is assigned the theme (rather than the agent) role in the sentence and that the sentence is transitive. Both these boost the RR posterior resulting in the situation shown in Figure 22.

4. Thus after the "by" phrase, there is a **flip**. The previously top ranked interpretation (MC) is now second ranked and the previously second ranked interpretation (RR) is now the top ranked interpretation. Our model predicts that such a *flip* correlates with an increased reading time effect.

Figure 23 shows the change in the posterior probability for the MC RR interpretations for the Good Patient (GP) cases. The data is averaged over the 40 verb/GP sentences in the McRae data. The results show the following effects.

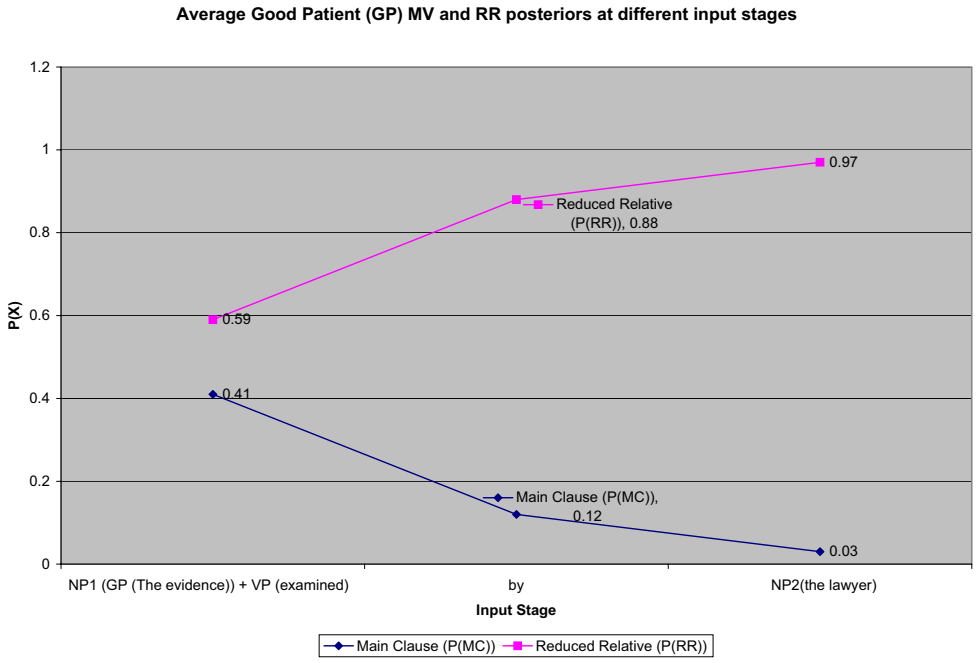

Figure 23: P(MC) and P(RR) for the ambiguous region for the Good Patient (GP) case

1. At the initial NP phrase the MC interpretation is more than twice as likely as the RR interpretation. This reflects the $SCFG$ structural probabilities prior to seeing the verb.

2. At the verb boundary phrase the RR interpretation is on average slightly higher ($P(MC) = .59$). This reflects the fact that both transitive bias of the verbs as well as the better thematic fit of the initial NP to the them (favouring the RR interpretation).

3. After the "by" phrase, things change even more. Now we notice that the RR posterior is almost five times as high as the MC posterior. This reflects the RR bias at 30 the by phrase, where there is now a much higher probability that the initial NP is assigned the theme (consistent with the previous assignment) role in the sentence and that the sentence is transitive. Both these boost the RR posterior resulting in the situation shown in Figure 23.

4. Thus after the "by" phrase, the previously top ranked interpretation (RR) continues to be the top ranked interpretation receiving more syntactic and thematic/semantic support while the second ranked interpretation (MC) continues to be second ranked. Hence in the good patient (GP) sentences, there is **no flip**. Thus our model does not predict increased reading time effects for GP sentences.

In summary, Figure 22 and Figure 23 show how the human reading time reduction effects (reduced compared to control sentences) increase for Good Agents (GA) but decrease for Good Patients in the ambiguous region. This is consistent with the reading time effect in the data in Figure 1. Our model predicts this larger effect from the fact that the most probable interpretation for the Good Agent case **flips** from the MC to the RR interpretation in this region. No such flip occurs for the Good Patient (GP) case. In Figure 23, we see that the GP results already have the $\frac{MC}{RR}$ ratio less than one (the RR interpretation is superior) while a flip occurs for the GA sentences (from the initial state where $\frac{MC}{RR} > 1$ to the final state where $\frac{MC}{RR} < 1$. This finding is fairly robust (85% of GA examples) and directly predicts reading time difficulties. In contrast, all (100%) of GP examples show **no flip**, and no reading time difficulty is predicted for these examples.

5.2.4 Quantifying the reading time effect

While a **flip** predicts increased reading time, can we quantify the magnitude of the effect? As we discussed earlier, we know the absolute value of reading time is dependent on many factors, including word length, grapheme or phoneme probability and transition probability, word imageability, punctuation, the specific location of the phrase on the text, and a wide variety of individual differences in working memory, reading speed, and other factors. Our model of course has no way to model these factors. Our model instead will attempt to capture only differences in relative reading time.

Equation (61), repeated here as Equation (75), shows our prediction for relative reading time differences:

$$\text{ProcessingTime}_{\text{Expectation}} \approx -\delta(I_{i^t}) \tag{75}$$

More specifically, since any sentence has multiple interpretations, the total magnitude of the change in the probability mass from word $t-1$ to word t is

$$\delta(t) = \sum_i \frac{P(I_i^t)}{P(I_i^{t-1})} \tag{76}$$

Thus our model predicts that the magnitude of the total change in conditional probabilities (summed over all interpretations) should correlate with changes in reading time.

Figure 25 and Figure 26 show the reading time effects predicted by our model compared to the reading time effects observed by McRae *et al.* (1998). Both sets of values are re-normalized as described slightly later in this section, and so the graph shows only the correlation between our probabilistic predictions and reading time, by normalizing both values and showing them on the same graph.

5.2.5 $\delta(t)$ alone is insufficient: *Flip* has a specific effect

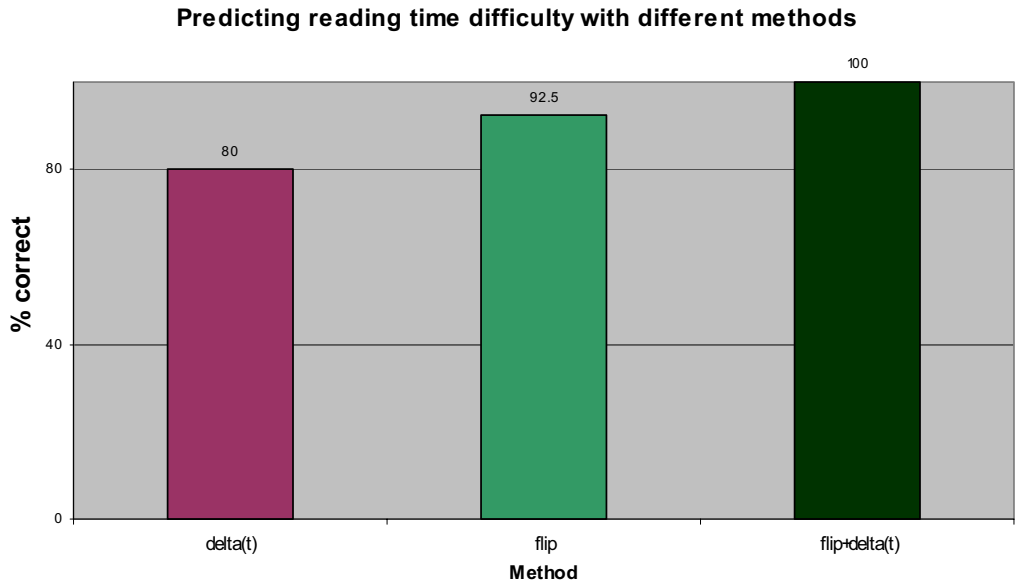

Figure 24: Reading time difficulty predictions from delta(t) alone, flip alone and the combination of flip and delta(t) (flip was weighted 2.0)

Figure 25 and Figure 26 show that the correlation between our probabilistic predictions and reading time seems reasonable; on average we do indeed predict the general locus of reading time increases.

But when we look at the correlation with individual sentences, we see a problem. We computed the $\delta(t)$ values for the ambiguous region in each of the 40 sentences in the input. 100% (20/20) of the Good Patient (GP) items were classified correctly; in each case the value of $\delta(t)$ predicted no enhanced difficulty at reading the "by" phrase. But only 12 of the 20 Good Agent (GA) (60% of GA) cases were classified correctly; (i.e., the value of $\delta(t)$ only predicted enhanced reading times at the "by" phrase 60% of the time). Thus the overall classification accuracy using just the $\delta(t)$ values is 80% ($\frac{20+12}{40} = .8$). The first column of the bar chart in Figure 24 shows this result.

Thus the Expectation principle by itself is insufficient to explain the reading time results from (McRae *et al.*, 1998). Could the Attention principle predict the results? Suppose we made a simple prediction that any sentence with a flip caused a reading increase. Figure 24 shows the resulting reading time difficulty prediction when we just looked at which of the 40 sentences flipped at the "by" phrase. In this case, we found all 20 of the GP cases (100% of GP) did not show a flip, while 17 out of 20 of the GA cases (85% of GA) flipped and thus predicted reading time difficulty. Thus just looking at the cases of flip alone, we had an overall classification accuracy of 92.5% ($\frac{17+20}{40} = .925$). The second column of the bar chart in Figure 24 shows this result.

Neither of these models accounts for all of the data. We instead applied the model we introduced in the introduction, repeated here as (77).

$$\text{ProcessingTime}_{\text{Reordering}} \approx$$
$$= \begin{cases} -w_{flip} \times \delta(I_{i^t}), & \text{if } P^*(I^t) \neq P^*(I^{t-1} \\ -\delta(I_{i^t}), & \text{if } P^*(I^t) = P^*(I^{t-1} \end{cases}$$

This equation combines the expectation and attention principles by making a combined prediction for reading time increases, using a single weight $w_{reorder}$. Figure 24 shows the reading time difficulty predictions when we set $w_{reorder}$ to 2, which corresponds to weighing data points that exhibited a flip as contributing double to the reading time increase. This combined model of flip and $\delta(t)$ explains all the data point, with an overall classification accuracy of 100% ($\frac{20+20}{40} = 1$). The third column of the bar chart in Figure 24 shows this result.

We experimented with a few parameter settings for weighting the flip data. The best results were obtained for values in range (1.2 3) ($1.2 < w_{flip} < 3$). The results shown here use a value roughly in the middle of that range (2). While our model of this data thus suggests that a flip causes additional reading time beyond the normal expectation violation based on $\delta(t)$, we will need more data to properly quantify this enhanced reading time effect. The next section details our reading time results with the combination of $\delta(t)$ and flip metric ($w_{flip} = 2$).

5.2.6 Reading time results with the combined ($\delta(t)$ + flip) metric

For our final set of reading time models, we attempted to model the difference between the reduced ("The cop examined") versus unreduced ("The cop that was examined") reading times at each of the stages in the input. Again, since our model does not predict absolute values of reading time, we compared the magnitude of the change in conditional probabilities with the percentage of the reading time effect at a particular input stage. To compute the scaled reading time effect in the data, we measured the percentage of reading time effects at a given stage from the McRae *et al.* (1998) data (Figure 5 from McRae *et al.* (1998)). For example, if the total reading time effect was 100 ms (over all stages) and the the particular stage (say at the verb) was 20 ms. then the percentage contribution of that stage was computed to be $\frac{20}{100} = .2$.

To compute the reading time effect as predicted by the model, we calculated the percentage of the total change in conditional probabilities weighted by the flip weight (of 2) (summed over all interpretations at all stages) contributed by a particular input stage. For instance, the total change was 10, and the change contributed by a particular stage was 2, then the model predicts an effect of magnitude $\frac{2}{10} = .2$ for that stage.

Figure 25 shows the reading time effect for the Good Agent (GA) sentences predicted by the model compared to the McRae reading time data. The results are averaged over the 40 verb/GA sentence pairs in the McRae data. In each case, the posterior probability was computed for the reduced (NP verb) at different stages of the input. The total effect was summed over the ambiguous region of the input (Initial NP + verb, PP, Second NP). The scaled effects were then

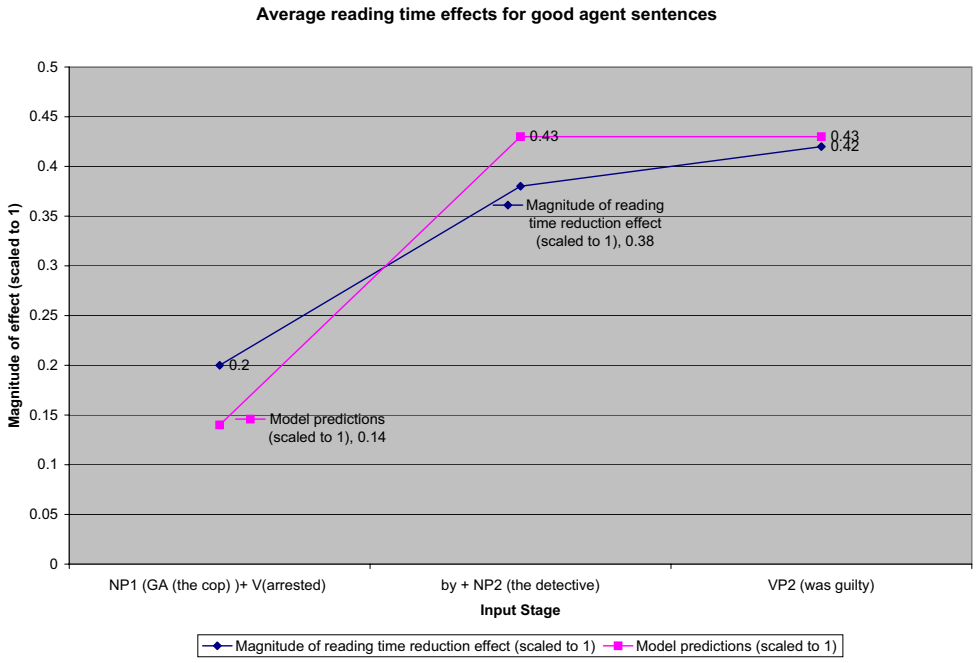

Figure 25: Reading time effect for GA sentences in the ambiguous region comparing the model predictions to data from McRae

computed quantifying the magnitude of an individual stage. The reading time effect from the data was computed as described above and scaled to 1.

In Figure 25 we see a fairly close match between the scaled human performance data and the model predictions of the reading time effects. In this case, the following facts emerge from Figure 25.

1. At the verb boundary (X-ed) we find on average a low reading time effect Thus the model is more likely to show an reduced reading time (compared to the unreduced case) at the verb boundary for the good agent case. In the model roughly .14 of the magnitude of the reading time effect was at this input stage. This is consistent with the human performance data in McRae *et al.* (1998). The scaled (to 1) value for the human data shows an effect of around .2.

2. After the "by" phrase, we find on average an *enhanced* reading time effect. Thus the model is more likely to show an enhanced reading time (compared to the unreduced case) after the "by" phrase boundary for the good agent case. In the model roughly .43 of the magnitude of the reading time effect was at this input stage. This is consistent with the human performance data in McRae *et al.* (1998). The scaled (to 1) value for the human data shows an effect of around .38.

3. Consistent with the previous observation, at the second NP phrase, we find on average an *enhanced* reading time effect. Thus the model is more likely to show an enhanced reading time (compared to the unreduced case) after the second NP phrase boundary for the good agent case. In the model roughly .43 of the magnitude of the reading time effect was at this input stage. This is consistent with the human performance data in McRae *et al.* (1998). The scaled (to 1) value for the human data shows an effect of around .42.

Figure 26 shows the reading time effect for the Good Agent (GA) sentences predicted by the model compared to the McRae reading time data. The results are averaged over the 40 verb/GA sentence pairs in the McRae data. In

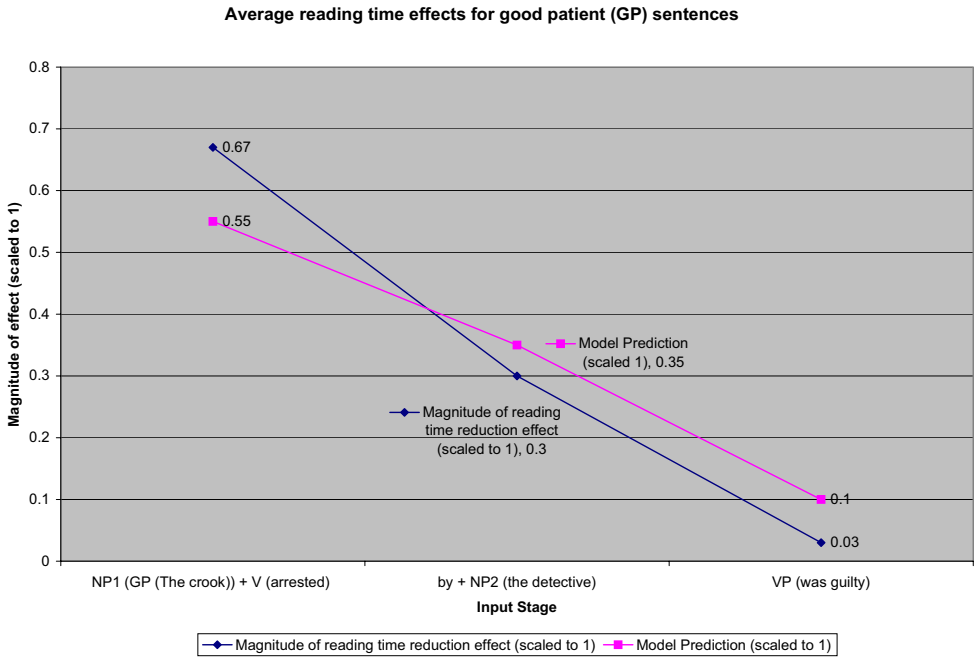

Figure 26: Reading time effect for GP sentences in the ambiguous region comparing the model predictions to data from McRae

each case, the posterior probability was computed for the reduced (NP verb) with the unreduced (NP that was Verb) at different stages of the input. The total effect was summed over the ambiguous region of the input (Initial NP + verb, PP, Second NP). The scaled effects were then computed quantifying the magnitude of an individual stage. The reading time effect from the data was computed as described above and scaled to 1.

Again we see a fairly close match between the scaled human performance data and the model predictions of the reading time effects. In this case, the following facts emerge from Figure 26.

1. At the verb boundary (X-ed) we find on average the reading time effect is highest. Thus the model is more likely to show an **enhanced** reading time (compared to the unreduced case) at the verb boundary for the good patient case. In the model roughly .55 of the magnitude of the reading time effect was at this input stage. This is consistent with the human performance data in McRae *et al.* (1998). The scaled (to 1) value for the human data shows an effect of around .67.

2. After the "by" phrase, we find on average a reduced reading time effect. Thus the model is less likely to show an enhanced reading time (compared to the unreduced case) after the "by" phrase boundary for the good patient case. In the model roughly .35 of the magnitude of the reading time effect was at this input stage. This is consistent with the human performance data in McRae *et al.* (1998). The scaled (to 1) value for the human data shows an effect of around .3.

3. After the second NP phrase, we find on average a much reduced reading time effect. Thus the model is much less likely to show an enhanced reading time (compared to the unreduced case) after the second NP phrase boundary for the good patient case. In the model roughly .1 of the magnitude of the reading time effect was at this input stage. This is consistent with the human performance data in McRae *et al.* (1998). The scaled (to 1) value for the human data shows an effect of around .03.

5.3 Summary of Results for Study One

In summary, our model for the McRae data shows the following effects.

1. As in McRae *et al.* (1998) the data shows that thematic fit clearly influenced the gated sentence completion task. The probabilistic account further captured the fact that at the *by* phrase, the posterior probability of producing an RR interpretation increased sharply, thematic fit and other factors influenced both the sharpness and the magnitude of the increase.

2. Our model predicts this larger reading time effect (see Figure 1) for the Good Agent(GA) sentences from the fact that the most probable interpretation for the Good Agent case *flips* from the MC to the RR interpretation in this region. No such flip occurs for the Good Patient (GP) case. In Figure 23, we see that the GP results already have the MC/RR ratio less than one (the RR interpretation is superior) while a flip occurs for the GA sentences (from the initial state where MC/RR > 1 to the final state where MC/RR < 1). This finding is fairly robust (85% of GA examples) and directly predicts reading time difficulties.

3. Our model shows that the magnitude of the reading time effect is correlated to **both** a) the magnitude of change of the conditional probabilities and b) the flip effect. The size of reading time effect at any stage is thus predictable from modulating the the change in conditional probabilities (summed over all interpretations) whenever there is a flip at that stage. Intuitively, this is consistent with the flip theory and offers a quantitative model for predicting reading time effects.

6 Study Two: The DO/SC Ambiguity and Pickering *et al.* (2000)

The previous section showed that the Bayesian model, via the Expectation and Attention principles. was able to account for the variation in reading-time across the processing of the main clause/reduced relative ambiguity shown by (McRae *et al.*, 1998). As we saw in the introduction, however, no previous probabilistic model has been able to model both the (McRae *et al.*, 1998) results and the (Pickering *et al.*, 2000) results on the direct object/sentential complement (DO/SC) ambiguity. Besides the importance of testing our model on more than one class of ambiguity, the (Pickering *et al.*, 2000) study is important also because their results have been interpreted as a direct argument against frequency-based models of any sort. Accounting for these results is thus a crucial test of our model.

6.1 The data

Recall that Pickering *et al.* (2000) studied DO/SC ambiguities in which the post-verbal noun was an implausible direct object of the verb, like *exercises* below:

(77) The young athlete realized her potential one day might make her a word-class sprinter.

(78) The young athlete realized her exercises one day might make her a word-class sprinter.

Pickering *et al.* (2000) showed that reading time was delayed on the phrase *one day* after the implausible direct object *her exercises* but not after the plausible direct object *her potential*. In other words, reading time on the phrase *one day* was higher in 78 than after 77. Since the verbs (like *realize*) were S-bias verbs, as shown in norming studies, this implies that a further reduction in plausibility of the less-plausible interpretation caused a reading-time increase.

Their materials were based on 6 verbs (*admitted, examined, decided, hinted, implied, and pretended*) and 16 sentence-pair items such as the one above. In order to norm the verbs, (Pickering *et al.*, 2000) had participants complete sentences with them, both in isolation and with a subject noun phrase. In the second test only subject-verb completions that produced twice as many SC as DO completions were used in the reading time study. The plausibility norming study asked subjects to assign a number from 0 to 7 for various postverbal noun phrases (such as potential or exercises above). For plausible NPs (*potential* in "The young athlete realized her potential."), the lower bound was 5 or higher (out of 7). For implausible NPs (*exercises* in "The young athlete realized her exercises"), the NP was used if the plausibility rating was 2 or lower (out of 7).

6.2 The model

Our model predicts two ways that a probabilistic model can explain increased reading time: a low probability assignment to the next word (equal to a large change in probability mass $\delta(t)$), or a 'flip', i.e. a demotion of the best interpretation.

The parameters of the model were set from the norming data computed by Pickering *et al.* (2000), consisting of:

$$\begin{aligned} P(SC|verb) \\ P(DO|verb) \\ P(SC|initialNP, verb) \\ P(DO|initialNP, verb) \end{aligned} \tag{79}$$

For example for the sentence prefix "The young athlete realized her", we used the following probabilities from Pickering *et al.* (2000). [9]

Parameter	Value
$(P(SC)\|V = realized)$	.35
$(P(DO)\|V = realized)$	0.25
$(P(SC)\|VP = realized, [_{NP}her \ldots], InitialNP = The, young, athlete)$	.8
$(P(DO)\|VP = realized, [_{NP}her \ldots], InitialNP = The, young, athlete)$	0.2

Table 5: The verb bias and the subcategorization probabilities for the sentence fragment "The young athlete realized her.."

In order to test the model, we thus need to measure the probability assigned to the interpretation before and after the implausible direct object *exercises*. Since our model rebuilds the Bayes net after each word, each probability would be generated by a slightly different net.

Figure 27 shows the structure of the Bayes net just after the direct object has been read. The top row shows the syntactic and lexical/thematic networks for the SC interpretation, while the bottom row shows these networks instantiated for the DO interpretation. The probability for each interpretation (DO or S) is computed given the sentence so far. The NOISY-AND combination function is applied to combine the lexical/thematic and syntactic support to arrive at the overall posterior probability of an interpretation at a particular stage of the input.

Recall that our model predicts that reading time is proportional to to change in the probability mass from word $t-1$ to word t, or

$$\delta(t) = \sum_i \frac{P(I_i^t)}{P(I_i^{t-1})} \tag{80}$$

6.3 Results

In this section we presently jointly the predictions of our model and the reading-time results of the Pickering *et al.* (2000) study.

We begin with an illustrative example, walking through the probabilities our model assigns to the following two sentences:

1. **Plausible Object**: The athlete realized her potential one day might make her a world class sprinter.
2. **Implausible Object**: The athlete realized her exercises one day might make her a world class sprinter.

Table 6 shows the $\delta(t)$ computed by our model after each region of both input sentences.

For the implausible object condition, the posteriors behaved in the following manner:

[9] Note the verb bias data don't sum up to 1, since there are other possible sentence completions for the verb.

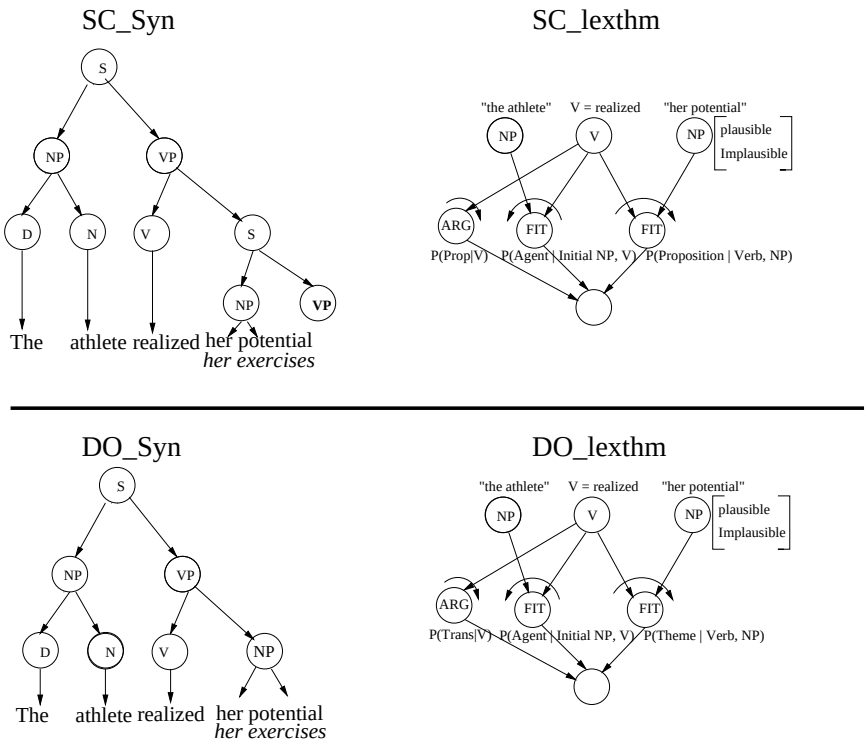

Figure 27: SC vs Object analysis: modeling the Pickering *et al.* (2000) data.

Realized	Matrix Verb	Noun	Post-noun	Modal verb
$(\delta(t)\|exercises)$	1.4	3.2	1.6	1.5
$(\delta(t)\|potential)$	1.0	1.5	1.4	3.3

Table 6: $\delta(t)$ results of the model on example sentences "The athlete realized her *exercises* one day might make her a world class sprinter" (Implausible object), and "The athlete realized her *potential* one day might make her a world class sprinter" (Plausible Object).

- After reading the matrix verb (*realized*), $P(I_{sc} > P(I_{do})$, so the model prefers the SC reading (these probabilities are not shown in Table 6). $\delta(t)$ (in Table 6 is 1.4, not particularly high.
- After reading the pronoun *realized her*, $P(I_{sc} > P(I_{do})$, so the model continues to prefer the SC reading.
- After seeing the implausible (*exercises*) direct object, there is a high drop in the posterior for the object interpretation ($\delta(t)$ jumps from 1.4 to 3.2). This is a large change in probability; the word *exercises* is unexpected and the interpreter is surprised. Our model thus predicts a reading time increase at this point.
- No large changes in probability mass happen in later words.

For the plausible object (*potential*) case, the posteriors behaved in the following manner

- After seeing the matrix verb (*realized*), $P(I_{sc} > P(I_{do})$, so the model prefers the SC reading. There is no change in probability mass
- At the preposition *realized her*), $P(I_{sc} > P(I_{do})$, so the model continues to prefer the SC reading.
- After seeing the plausible (*potential*) direct object, there is a small drop in the posterior for the object interpretation. $\delta(t)$ is 1.5, not particularly high, and our model not predict any reading time increase at this point.

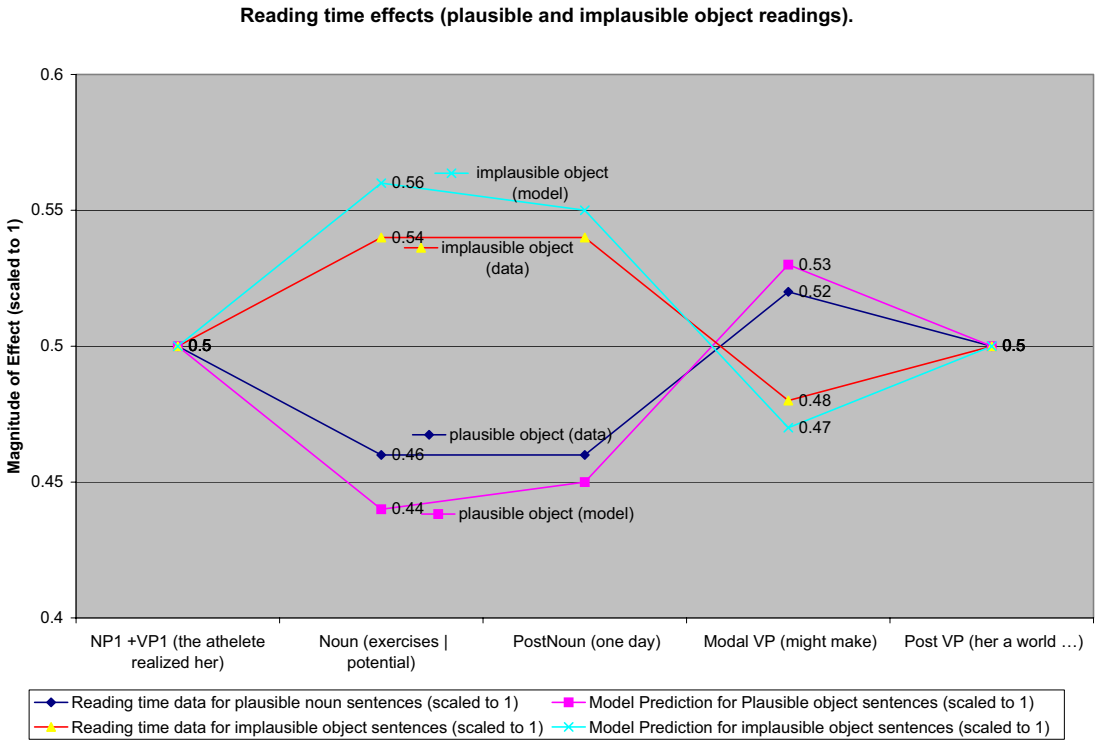

Figure 28: Overall Predictions for Pickering data. Shown are the model and scaled (to 1) human reading time effects (please see text for the details) for both plausible object and implausible object sentences.

- At the post-noun *one day* region, $\delta(t)$ is 1.5, not particularly high, and our model not predict any reading time increase at this point.
- At the disambiguating modal verb (*might*), the modal has a low probability, i.e. there is a large change in probability mass, $\delta(t)$ increases from 1.4 to 3.3. Our model registers an expectation violation and predicts an increase in reading time at this stage.

The walk-through above shows the qualitative results of our model; some regions show a larger probability mass change $\delta(t)$. Do these areas correspond to regions of longer reading time? To answer this question, Figure 28 compares the overall reading time effects predicted by our model to those observed by Pickering *et al.* (2000). The results are averaged over the six verbs (admitted, decided, hinted, implied, pretended and realized) and sixteen sentences for the two conditions (implausible and plausible object) in the Pickering *et al.* (2000) experiment.

Because this figure is somewhat difficult to read, we break this figure down in Figure 29 and Figure 28. Figure 29 compares the reading time effects predicted by our model for plausible object sentences (ex. The athlete realized her potential..) to those observed by Pickering *et al.* (2000). Figure 30 compares the reading time effects predicted by our model for plausible object sentences (ex. The athlete realized her exercises..) to those observed by Pickering *et al.* (2000).

How were these figures generated? In Figure 28, the human data is obtained from Pickering *et al.* (2000) (Table 3, pp. 456). Their reading time data was measured at four different input stages for the two conditions, plausible object (*potential*) and implausible object (*exercises*). The reading time was measured at 1) the noun (potential) , 2) the postnoun region (one day), 3) the disambiguating modal (might make) and 4) the postverb region (her a world class sprinter).

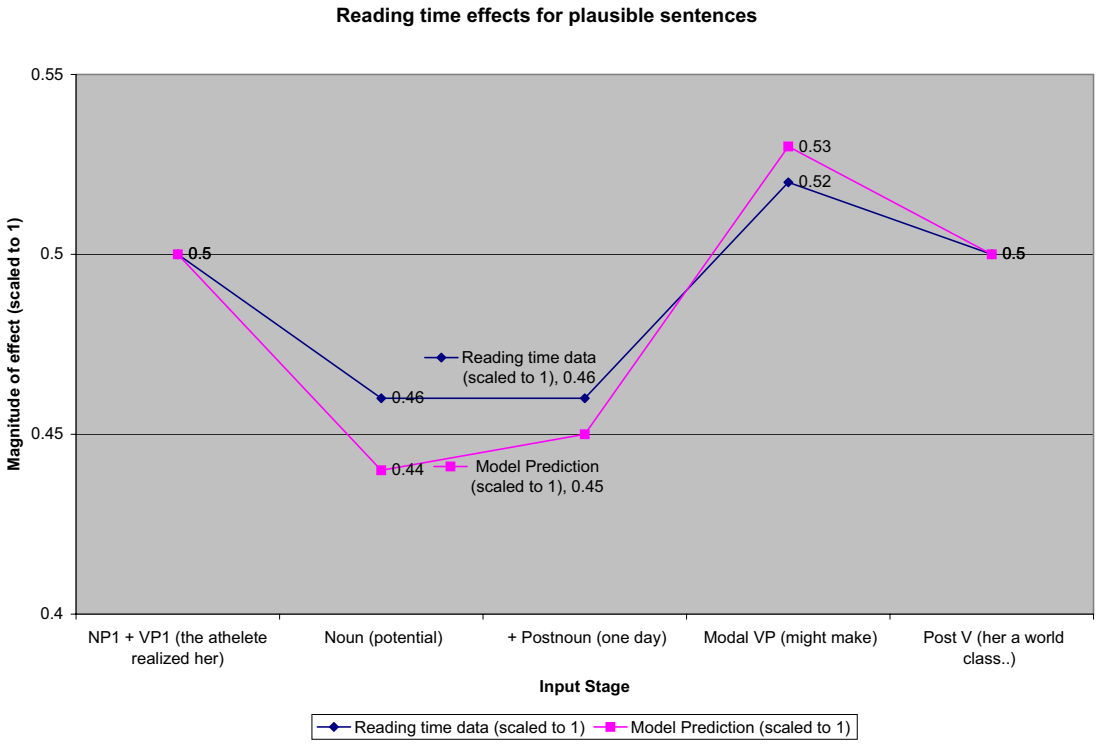

Figure 29: Predictions for plausible object sentences in the Pickering data. Shown are the model and scaled (to 1) human reading time effects (please see text for the details) for the plausible object sentences.

Our model does not make exact predictions about the absolute value of milliseconds of reading time from probabilities. Such a mapping would be possible in principle, but would require solving a number of problems that we simply don't currently have the data for, including choosing a function to map probability to time (log? cube-root?) as well as setting weight parameters. Instead, our goal is to show that the model makes the correct predictions about the *relative* magnitude of reading time increases.

In order to do this, we renormalize the human reading time data by computing the scaled (to 1) magnitude of the reading time contributions for the two conditions at each of the four input stages. For instance, at the Noun stage, Pickering *et al.* (2000) report a total reading time of 367 milliseconds for the plausible object condition, and a reading time of 430 milliseconds for the implausible object condition. So here the reading time effect for the plausible case is $\frac{367}{367+430} = .46$ and $\frac{430}{367+430} = .54$ for the implausible case. Figure 28 shows the value of the reading time effect computed in this way for all the four stages and for the two conditions in the Pickering *et al.* (2000) data.

For the model, we computed the total change in the posteriors for the two conditions between the various input stages. We fixed the initial baseline for the change computation at the pre-noun (The athlete realized her) stage. Thus the first change in posterior compares the value of the change in posteriors for the two conditions between the pre-noun and the noun stage. For the model, we computed the change in posteriors up to the disambiguating modal stage.[10]. As in the case of the human data we measured the relative contribution of the two conditions (scaled to 1).

Figure 28 shows the basic *crossover* result found by Pickering *et al.* (2000). Participants exhibited greater reading time difficulty starting at the noun boundary and continuing to the post-noun region for implausible sentences. However at the disambiguating modal verb the reading time was larger for plausible compared to the implausible

[10]We did not compute the change in posteriors for the post-verb stage since it was not relevant to the Pickering *et al.* (2000) results

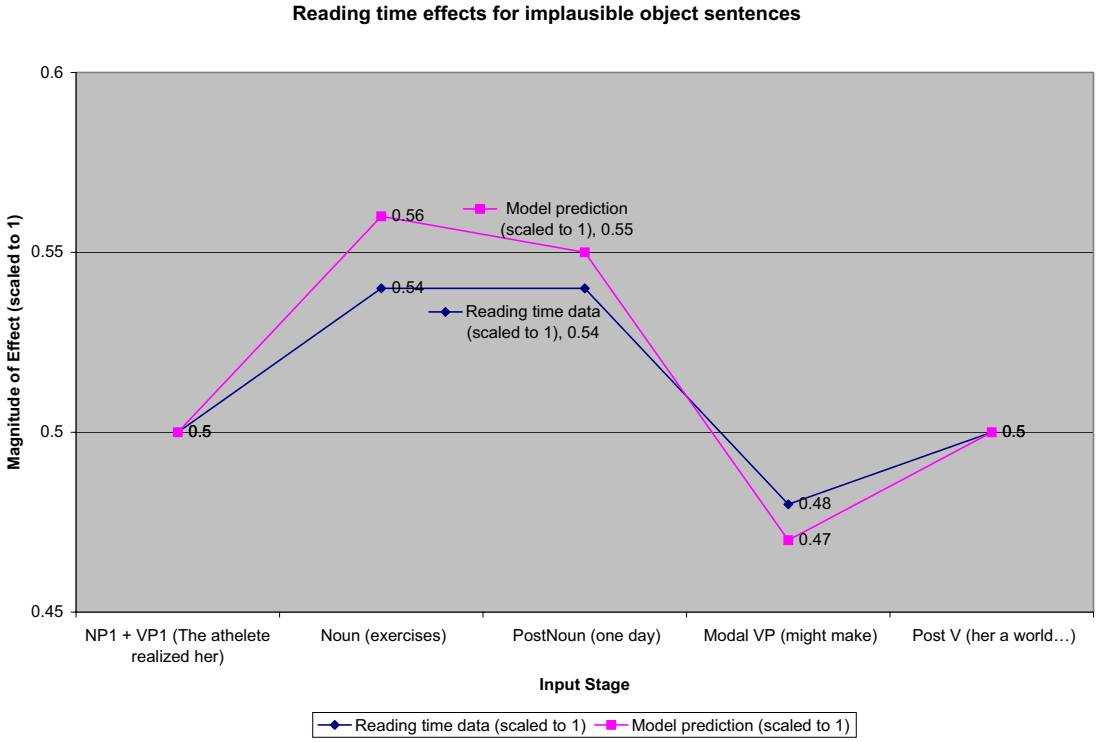

Figure 30: Predictions for implausible object sentences in the Pickering data. Shown are the model and scaled (to 1) human reading time effects (please see text for the details) for the implausible object sentences.

readings.

Thus as shown in Figure 28 our model predicts greater surprise and hence greater reading time effects at the initial object NP for the implausible object but at the post disambiguating region (after *might* for the plausible object. This is consistent with the data in Pickering *et al.* (2000). Our quantification of expectation violation (or surprise) based on large changes in the sum of the posterior probabilities for the different competing interpretations is thus able to account for the *crossover effect* in the Pickering data using a structured probabilistic sentence processing model.

6.4 Discussion

In summary, our model predicts two effects for the data in Pickering *et al.* (2000). First, the model predicts a greater reading time at the direct object noun for the implausible object than the plausible object, since the conditional probability of the noun is quite low for the implausible reading. In contrast, for the plausible object case, there is no significant change in the posteriors, since the noun is approximately equally likely under both interpretations, so the sum of the posteriors does not change much between the pre-noun and the noun boundary. This is consistent with the human data shown in Figure 30

Second, by contrast, the model predicts a greater reading time at the modal verb *might* for the plausible object case. This is because reading the verb causes a significant change in sum of the posterior probabilities of the two interpretations, since the verb is extremely unlikely for the direct object reading. In contrast, for the implausible object case, there is no significant change in the posteriors, since the object reading is already low, and the sum of the posteriors does not change much between the post-noun and the verb boundary. This prediction is consistent with the human data shown in Figure 29.

7 General Discussion

This paper introduced a probabilistic model for human sentence processing, which computes probabilities incrementally, integrates probabilistic versions of linguistic knowledge online, and makes predictions about parse preference and reading time. Our predictions, from the principles of Expectation and Attention, were that the probability of upcoming words is one key predictor of reading time, and that demotion of the top-ranked interpretation is another. We tested the model against behavioral experiments studying the disambiguation of the main clause/reduced relative and the direct object/sentential complement ambiguity. No previous probabilistic model has been able to model both of these classes of results. We showed that our model is able to explain the interpretation preferences as well the relative increases and decreases in reading time from each experiment.

In understanding the implications of any computational model of human cognition, it is important to distinguish potential insights into human language understanding from mere implementation details and other assumptions made for purely practical reasons. We think our model does offer some high-level insights into human language processing. First, we believe that human language processing is inherently probabilistic. Second, we believe that human language processing makes use of a variety of rich sources of linguistic knowledge at many levels. Third, since human language processing is on-line, any such model of this probabilistic process must also be able to model this dynamic process. We believe our Bayesian model provides one vision of how these three constraints (probabilistic computation, incremental update, combination of structured and probabilistic knowledge) can be viewed.

The relationship between probabilistic models and reading time expressed in the Expectation and Attention principles can also be viewed as a high-level insight that may carry over into other classes of models.

Beyond the high-level claims of our model, other aspects of our work may point a direction for integration with other current models. One of the problems with the competition model was its lack of a motivated way of modeling language structure, the class of possible constraints, and the weights on evidence combination. The Bayesian model provides a way of answering all these questions. Thus a hybrid model may be able to capture aspects of both models, perhaps making predictions about reading time effects due to either competition, expectation, or attention.

The structured probabilistic aspects of our model may also have a role in modeling language production. In a class of models dating back to Schuchardt, linguistics have argued that human lexical production is sensitive to the predictability of words. A series of experiments by the second author and colleagues (Gregory, Raymond, Bell, Fosler-Lussier, & Jurafsky, 1999; Jurafsky, Bell, Gregory, & Raymond, 2001; Jurafsky, Bell, & Girand, 2002) have shown that this predictability can be measured probabilistically, and proposed that the reduction or shortening in the surface form of words is proportional to the conditional probability of the word. But that work has so far not proposed a model of how various probabilities combine to predict the posterior probability of a word. We think our Bayesian model could help show how this is done.

8 Appendix A: Propagation in Bracketed SCFG trees

These are the derivations of the belief propagation rules for bracketed SCFG trees. Without loss of generality, we assume that the SCFG grammar is in the Chmosky Normal Form (CNF). Thus any node in the Parse tree has at most two children. Clearly in an SCFG production $x \rightarrow yz$ the specific non-terminals y and z are not independent given x unlike in standard Bayes nets.

We can modify the propagation rules for Bayes nets which have a tree structure to reflect this dependence. We will use the notation from (Pearl 1988) including the convention that e^+_X and e^-_X denote the causal and diagnostic evidence, respectively, relative to a node X. We now compute the diagnostic (bottom-up) and causal (top-down) support some nodes X, Y, and Z in the SCFG parse tree. By convention, we will use the lower case x to represent the instantiation of the variable X, y to represent the instantiation of node Y and z to represent the instantiation of node Z. In this case x, y, and z range over the non-terminals in the grammar.

Diagnostic support:

$$\begin{aligned}
\lambda(x) &= P(e^-_X|x) && (81)\\
&= \sum_{y,z} P(e^-_X, y, z|x) && (82)\\
&= \sum_{y,z} P(e^-_X|y, z, x)P(y, z|x) && (83)
\end{aligned}$$

$$= \sum_{y,z} P(e_X^-|y,z)P(y,z|x) \tag{84}$$

(because Y, Z separates X from e_X^-) (85)

$$= \sum_{y,z} P(e_Y^- \cup e_Z^-|y,z)P(y,z|x) \tag{86}$$

$$= \sum_{y,z} P(e_Y^-|y)P(e_Z^-|z)P(y,z|x) \tag{87}$$

(because Y, Z separate e_Y^-, e_Z^- from each other) (88)

$$= \sum_{y,z} \lambda(y)\lambda(z)P(y,z|x) \tag{89}$$

Causal support:

$$\pi(x) = P(x|e_X^+) \tag{90}$$

$$= \sum_{u,v} P(x,u,v|e_X^+) \tag{91}$$

$$= \sum_{u,v} P(x|u,v,e_X^+)P(u,v|e_X^+) \tag{92}$$

$$= \sum_{u,v} P(x|u,v)P(u,v|e_X^+) \tag{93}$$

(because U,V separates X from e_X^+) (94)

$$= \frac{1}{P(e_X^+)} \sum_{u,v} P(x|u,v)P(u,v,e_X^+) \tag{95}$$

The second term in the summation can be expanded as follows:

$$P(u,v,e_X^+) = P(u,v,e_U^+ \cup e_V^-) \tag{96}$$

$$= P(u,e_U^+)P(v,e_V^-|u,e_U^+) \tag{97}$$

$$= P(u,e_U^+)P(v,e_V^-|u) \tag{98}$$

(because U separates V from e_U^+) (99)

$$= P(u|e_U^+)P(e_U^+)P(e_V^-|v,u)P(v|u) \tag{100}$$

$$= P(u|e_U^+)P(e_U^+)P(e_V^-|v)P(v|u) \tag{101}$$

(because V separates U from e_V^-) (102)

$$= \pi(u)P(e_U^+)\lambda(v)P(v|u) \tag{103}$$

Substituting back into the equation for π,

$$\pi(x) = \frac{P(e_U^+)}{P(e_X^+)} \sum_{u,v} \pi(u)\lambda(v)P(x|u,v)P(v|u)$$

$$= \frac{P(e_U^+)}{P(e_X^+)} \sum_{u,v} \pi(u)\lambda(v)P(x,v|u) \tag{104}$$

The normalizing constant $\frac{P(e_U^+)}{P(e_X^+)} = \frac{1}{P(e_V^-|e_U^+)}$ can be computed implicitly by scaling the $\pi(x)$ to sum to unity.

Finally, it can be seen how the outer probabilities arise naturally by using the propagation scheme for π without normalization:

$$f(y) = \sum_{x,z} f(x)\lambda(z)P(y,z|x) \tag{105}$$

Simple substitution in (104) shows that the fixed point for the functional equation (105) is $f(x) = \pi(x)P(e_X^+)$, the generalized outer probability.

9 Appendix B: Parameters for Experiment 1

Verb	Intrans		GoodAgt		GoodPat	PastPart	SimplePast
		A	P	A	P		
accuse	0.279452	.67	.33	.67	.46	.53	
arrest	0.530612	.81	.19	.17	.83	.81	.19
capture	0.317016	.75	.25	.34	.66	.63	.37
carry	0.329179	.81	.19	.16	.84	.77	.23
chase	0.380952	.72	.28	.36	.64	.49	.50
convict	0.533333	.67	.33	.19	.81	.84	.16
cure	0.440678	.64	.36	.19	.81	.80	.20
devour	0.396825	.61	.39	.37	.63	.49	.51
dismiss	0.275176	.76	.24	.20	.80	.61	.39
entertain	0.399160	.76	.24	.20	.80	.54	.45
evaluate	0.344086	.62	.38	.37	.63	.87	.13
examine	0.310172	.74	.26	.32	.68	.63	.37
execute	0.392857	.57	.43	.40	.60	.78	.22
fire	0.553366	.72	.28	.23	.77	.50	.50
frighten	0.156627	.78	.22	.27	.73	.60	.40
grade	0.714286	.73	.27	.25	.75	.84	.16
hire	0.381271	.70	.30	.19	.81	.57	.43
hypnotize	0.666667	.74	.26	.20	.80	.76	.34
instruct	0.320388	.76	.24	.23	.77	.58	.42
interrogate	0.565217	.80	.20	.23	.77	.75	.25
interview	0.463722	.72	.28	.28	.72	.65	.35
investigate	0.345679	.74	.26	.36	.64	.76	.24
invite	0.161580	.60	.40	.23	.77	.68	.32
kick	0.560680	.64	.36	.23	.77	.30	.70
lecture	0.800000	.72	.28	.26	.74	.34	.66
lift	0.467054	.71	.29	.30	.70	.42	.58
punish	0.282051	.68	.32	.20	.80	.85	.15
question	0.557452	.69	.31	.23	.77	.61	.39
recognize	0.582450	.61	.39	.36	.64	.71	.29
rescue	0.358209	.78	.22	.21	.79	.71	.29
search	0.701571	.82	.18	.21	.79	.38	.62
sentence	0.289157	.84	.16	.19	.81	.80	.20
serve	0.711992	.73	.27	.18	.82	.69	.31
shoot	0.608943	.71	.29	.14	.86	.63	.37
slaughter	0.406250	.81	.19	.13	.87	.70	.30
study	0.479351	.65	.35	.41	.59	.62	.38
teach	0.425798	.72	.28	.20	.80	.26	.74
terrorize	0.478261	.75	.25	.18	.82	.98	.02
torture	0.630435	.78	.22	.19	.81	.75	.25
worship	0.763975	.62	.38	.18	.82	.45	.55

10 Appendix C: Results for Experiment 1

Verb	(at Initial NP)			(at verb-ed)			(at by)			(at the)			(at agent NP)		
	P(M)P(R)P(M)/P(R)			L(M)L(R)P(M)/P(R)			L(M)L(R)P(M)/P(R)			L(M)L(R)P(M)/P(R)			L(M)L(R)P(M)/P(R)		
accuse															
GA	.667	.333	2.00	.355	.156	2.282	.071	.125	.568	.009	.109	.082	.002	.090	.022
GP	.327	.673	0.49	.174	.314	0.553	.035	.251	.139	.004	.122	.033	.001	.100	.010
arrest															
GA	.807	.193	4.19	.170	.156	1.083	.034	.125	.272			.039			.005
GP	.169	.831	0.20	.032	.671	0.048			.011			.002			.0002
capture															
GA	.750	.250	3.00	.278	.158	1.762			.440			.063			.011
GP	.338	.662	0.51	.125	.416	0.301			.075			.011			.002
carry															
GA	.810	.190	4.25	.189	.146	1.295			.324			.046			.008
GP	.162	.838	0.19	.038	.643	0.059			.014			.002			.0002
chase															
GA	.667	.333	2.00	.355	.156	2.282			.568			.082			.022
GP	.327	.673	0.49	.174	.314	0.553			.139			.033			.010
convict															
GA	.846	.154	5.49	.138	.129	1.070			.268			.038			.007
GP	.192	.808	0.24	.031	.737	0.043			.010			.002			.0002
cure															
GA	.642	.358	1.79	.127	.288	0.440			.110			.016			.003
GP	.187	.813	0.23	.037	.653	0.056			.014			.002			.0003
devour															
GA	.614	.386	1.59	.311	.190	1.640			.410			.059			.010
GP	.375	.625	0.60	.191	.308	0.620			.155			.022			.004
dismiss															

```
 GA    .760 .240 3.16  .299 .145 2.060            .515             .074             .013
 GP    .203 .797 0.26  .080 .483 0.165            .041             .006             .001
entertain
 GA    .761 .239 3.19  .350 .130 2.69             .673             .096             .017
 GP    .202 .798 0.25  .093 .433 0.214            .054             .008             .001
evaluate
 GA    .618 .382 1.62  .083 .330 0.252            .063             .009             .001
 GP    .367 .633 0.58  .050 .548 0.091            .023             .003             .0005
examine
 GA    .744 .256 2.91  .278 .161 1.729            .432             .062             .011
 GP    .318 .682 0.47  .119 .590 0.201            .050             .007             .001
```

References

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum, Hillsdale, NJ.

Babyonyshev, M., & Gibson, E. (1999). The complexity of nested structures in Japanese. *Language*, *75*(3), 423–450.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 340–357.

Booth, T. L. (1969). Probabilistic representation of formal languages. In *IEEE Conference Record of the 1969 Tenth Annual Symposium on Switching and Automata Theory*, pp. 74–81.

Brants, T. (1999). Cascaded markov models. In *Proceedings of the 9th Conference of the European Chapter of the ACL (EACL-99)* Bergen, Norway. ACL.

Burgess, C., & Hollbach, S. C. (1988). A computational model of syntactic ambiguity as a lexical process. In *COGSCI-88*, pp. 263–269.

Burgess, C., & Lund, K. (1994). Multiple constraints in syntactic ambiguity resolution: A connectionist account of psycholinguistic data. In *COGSCI-94*, pp. 90–95 Atlanta, GA.

Chomsky, N. (1956). Three models for the description of language. *IRI Transactions on Information Theory*, *2*(3), 113–124.

Clifton, Jr., C., & Ferreira, F. (1987). Modularity in sentence comprehension. In Garfield, J. L. (Ed.), *Modularity in knowledge representation and natural-language understanding.*, pp. 277–290. MIT Press, Cambridge, MA.

Clifton, Jr., C., Frazier, L., & Connine, C. (1984). Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, *23*, 696–708.

Crocker, M. W., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, *29*(6), 647–669.

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

Fodor, J. A. (1983). *Modularity of Mind*. MIT Press, Cambridge, MA.

Ford, M., Bresnan, J., & Kaplan, R. M. (1982). A competence-based theory of syntactic closure. In Bresnan, J. (Ed.), *The Mental Representation of Grammatical Relations*, pp. 727–796. MIT Press, Cambridge, MA.

Forster, K., & Chambers, S. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627–635.

Frazier, L. (1987). Theories of sentence processing. In Garfield, J. L. (Ed.), *Modularity in knowledge representation and natural-language understanding*, pp. 291–307. MIT Press, Cambridge, MA.

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

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*, 291–295.

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.

Gibson, E. (1990a). Memory capacity and sentence processing. In *Proceedings of the 28th ACL* Pittsburgh, PA. ACL.

Gibson, E. (1990b). Recency preference and garden-path effects. In *COGSCI-90*, pp. 372–379 Cambridge, MA.

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

Glymour, C., & Cheng, P. W. (1998). Causal mechanism and probability: A normative approach. In Oaksford, M., & Chater, N. (Eds.), *Rational Models of Cognition*, pp. 296–313. Oxford University Press, Oxford.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *CLS-99*, pp. 151–166. University of Chicago, Chicago.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*, *28*, 267–283.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of NAACL-2001*, pp. 159–166.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, *29*, 296–305.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, *41*, 401–410.

Jelinek, F., & Lafferty, J. D. (1991). Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, *17*(3), 315–323.

Jennings, F., Randall, B., & TYler, L. K. (1997). Graded effects of verb subcategory preferences on parsing: Support for constraint-satisfaction models. *Language and Cognitive Processes*, *12*(4), 485–504.

Jespersen, O. (1922). *Language*. Henry Holt, New York.

Juliano, C., & Tanenhaus, M. K. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *COGSCI-93*, pp. 593–598 Boulder, CO.

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

Jurafsky, D., Bell, A., & Girand, C. (2002). The role of the lemma in form variation. In Warner, N., & Gussenhoven, C. (Eds.), *Papers in Laboratory Phonology 7*, pp. 1–34. Mouton de Gruyter, Berlin/New York.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee, J., & Hopper, P. (Eds.), *Frequency and the Emergence of Linguistic Structure*, pp. 229–254. Benjamins, Amsterdam.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 123–154.

Kim, A., Srinivas, B., & Trueswell, J. (2002). The convergence of lexicalist perspectives in psycholinguistics and computational linguistics. In Merlo, P., & Stevenson, S. (Eds.), *Sentence Processing and the lexicon: formal, computational, and experimental perspectives*, pp. 109–136. Benjamins, Amsterdam.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580–602.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182.

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*(2), 157–201.

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

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994b). Syntactic ambiguity resolution as lexical ambiguity resolution. In *Perspectives on Sentence Processing*, pp. 123–154. Erlbaum, Hillsdale, NJ.

McDonald, S., Shillcock, R., & Brew, C. (2001). Low-level predictive inference in reading: Using distributional statistics to predict eye movements. Poster presented at AMLaP-2001, Saarbruecken. September 20-22, 2001.

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

Miyake, A., Carpenter, P. A., & Just, M. A. (1994). A capacity approach to syntactic comprehension disorders: Making normal adults perform like aphasic patients. *Cognitive Neuropsychology, 11*, 671–717.

Moore, R., Appelt, D., Dowding, J., Gawron, J. M., & Moran, D. (1995). Combining linguistic and statistical knowledge sources in natural-language processing for ATIS. In *Proceedings of the January 1995 ARPA Spoken Language Systems Technology Workshop*, pp. 261–264 Austin, TX. Morgan Kaufmann.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, San Mateo, Ca.

Pearlmutter, N., Daugherty, K., MacDonald, M., & Seidenberg, M. (1994). Modeling the use of frequency and contextual biases in sentence processing. In *COGSCI-94*, pp. 699–704 Atlanta, GA.

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

Rehder, B. (1999). A causal model theory of categorization. In *COGSCI-99*, pp. 595–600 Vancouver, British Columbia.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9*, 487–494.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America, 35*, 200–206.

Schuchardt, H. (1885). *Über die Lautgesetze: Gegen die Junggrammatiker*. Robert Oppenheim, Berlin. Excerpted with English translation in Theo Vennemann and Terence H. Wilbur, (Eds.), *Schuchardt, the Neogrammarians, and the Transformational Theory of Phonological Change*, Athenaum Verlag, Frankfurt, 1972.

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science, 23*, 569–588.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1521–1543.

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

Spivey-Knowlton, M., Trueswell, J., & Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses.. *Canadian Journal of Experimental Psychology, 47*, 276–309.

Spivey-Knowlton, M. J. (1996). *Integration of visual and linguistic information: Human data and model simulations*. Ph.D. thesis, University of Rochester.

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics, 21*(2), 165–202.

Tabor, W., Juliano, C., & Tanenhaus, M. K. (1997). Parsing in a dynamical system. *Language and Cognitive Processes, 12*, 211–272.

Tabossi, P., Spivey-Knowlton, M., McRae, K., & Tanenhaus, M. K. (1994). Semantic effects on syntactic ambiguity resolution: Evidence for a constraint-based resolution process. In Umilta, C., & Moscovitch, M. (Eds.), *Attention and Performance XV*, pp. 589–615. Lawrence Erlbaum, Hillsdale, NJ.

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Hanna, J. E. (2000). Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. In Crocker, M. W., Pickering, M., & Clifton, C. (Eds.), *Architectures and Mechanims for Language Processing*, pp. 90–118.

Tanenhaus, M. K., Stowe, L. A., & Carlson, G. (1985). The interaction of lexical expectation and pragmatics in parsing filler-gap constructions. In *COGSCI-85*, pp. 361–365 Irvine, CA.

Tenenbaum, J. B. (2000). Bayesian modeling of human concept learning. In *Advances in Neural Information Processing Systems 11*.

Tenenbaum, J. B., & Griffiths, T. L. (2001a). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(3), 579–616.

Tenenbaum, J. B., & Griffiths, T. L. (2001b). The rational basis of representativeness. In *COGSCI-01* Edinburgh.

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. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In Clifton, Jr., C., Frazier, L., & Rayner, K. (Eds.), *Perspectives on Sentence Processing*, pp. 155–179. Lawrence Erlbaum, Hillsdale, NJ.

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

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*(3), 528–553.

Whaley, C. P. (1978). Word–nonword classification time. *Journal of Verbal Language and Verbal Behavior*, *17*, 143–154.

Wundt, W. (1900). *Völkerpsychologie: eine Untersuchung der Entwicklungsgesetze von Sprache, Mythus, und Sitte*. W. Engelmann, Leipzig. Band II: Die Sprache, Zweiter Teil.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, *15*, 1–95.