A noisy-channel model of rational human sentence comprehension under uncertain input

Roger Levy
Department of Linguistics
University of California – San Diego
9500 Gilman Drive #0108
La Jolla, CA 92093-0108
rlevy@ling.ucsd.edu *

Abstract

Language comprehension, as with all other cases of the extraction of meaningful structure from perceptual input, takes places under noisy conditions. If human language comprehension is a rational process in the sense of making use of all available information sources, then we might expect uncertainty at the level of word-level input to affect sentence-level comprehension. However, nearly all contemporary models of sentence comprehension assume *clean* input—that is, that the input to the sentence-level comprehension mechanism is a perfectly-formed, completely certain sequence of input tokens (words). This article presents a simple model of rational human sentence comprehension under noisy input, and uses the model to investigate some outstanding problems in the psycholinguistic literature for theories of rational human sentence comprehension. We argue that by explicitly accounting for input-level noise in sentence processing, our model provides solutions for these outstanding problems and broadens the scope of theories of human sentence comprehension as rational probabilistic inference.

* Part of this work has benefited from presentation at the 21st annual meeting of the CUNY Sentence Processing Conference in Chapel Hill, NC, 14 March 2008, and at a seminar at the Center for Research on Language, UC San Diego. I am grateful to Klinton Bicknell, Andy Kehler, and three anonymous reviewers for comments and suggestions, Cyril Allauzen for guidance regarding the OpenFST library, and to Mark Johnson, Mark-Jan Nederhof, and Noah Smith for discussion of renormalizing weighted CFGs.

1 Introduction

Considering the adversity of the conditions under which linguistic communication takes place in everyday life—ambiguity of the signal, environmental competition for our attention, speaker error, and so forth—it is perhaps remarkable that we are as successful at it as we are. Perhaps the leading explanation of this success is that (a) the linguistic signal is redundant, and (b) diverse information sources are generally available that can help us obtain infer the intended message (or something close enough) when comprehending an utterance (Tanenhaus et al., 1995; Altmann and Kamide, 1999; Genzel and Charniak, 2002, 2003; Aylett and Turk, 2004; Keller, 2004; Levy and Jaeger, 2007). Given the difficulty of this task coupled with the availability of redundancy and useful information sources, it would seem rational for all available information to be used to its fullest in sentence comprehension. This idea is either implicit or explicit in several interactivist theories of probabilistic language comprehension (Jurafsky, 1996; Hale, 2001; Narayanan and Jurafsky, 2002; Levy, 2008). However, these theories have implicitly assumed a partitioning of interactivity that distinguishes the *word* as a fundamental level of linguistic information processing: word recognition is an evidential process whose output is nonetheless a specific "winner-takes-all" sequence of words, which is in turn the input to an evidential sentence-comprehension process. It is theoretically possible that this partition is real and is an optimal solution to the problem of language comprehension under gross architectural constraints that favor modularity

(Fodor, 1983). On the other hand, it is also possible that this partition has been a theoretical convenience but that, in fact, evidence at the sub-word level plays an important role in sentence processing, and that sentence-level information can in turn affect word recognition. If the latter is the case, then the question arises of how we might model this type of information flow, and what consequences it might have for our understanding of human language comprehension. This article employs the well-understood formalisms of probabilistic context-free grammars (PCFGs) and weighted finite-state automata (wFSAs) to propose a novel yet simple noisy-channel probabilistic model of sentence comprehension under circumstances where there is uncertainty about word-level representations. Section 2 introduces this model. We use this new model to investigate two outstanding problems for the theory of rational sentence comprehension: one involving *global inference*—the beliefs that a human comprehender arrives at regarding the meaning of a sentence after reading it in its entirety (Section 3)—and one involving *incremental inference*—the beliefs that a comprehender forms and updates moment by moment while reading each part of it (Section 4). The common challenge posed by each of these problems is an apparent failure on the part of the comprehender to use information made available in one part of a sentence to rule out an interpretation of another part of the sentence that is inconsistent with this information. In each case, we will see that the introduction of uncertainty into the input representation, coupled with noisy-channel inference, provides a unified solution within a theory of rational comprehension.

2 Sentence comprehension under uncertain input

The use of generative probabilistic grammars for parsing is well understood (e.g., Charniak, 1997; Collins, 1999). The problem of using a probabilistic grammar G to find the "best parse" T for a known input string $\mathbf{w}$ is formulated as[1]

$$\arg\max_T P_G(T|\mathbf{w}) \quad \text{(I)}$$

but a *generative* grammar directly defines the joint distribution $P_G(T, \mathbf{w})$ rather than the conditional distribution. In this case, Bayes' rule is used to find the posterior:

$$P_G(T|\mathbf{w}) = \frac{P(T, \mathbf{w})}{P(\mathbf{w})} \quad \text{(II)}$$

$$\propto P(T, \mathbf{w}) \quad \text{(III)}$$

If the input string is unknown, the problem changes. Suppose we have some noisy evidence I that determines a probability distribution over input strings $P(\mathbf{w}|I)$. We can still use Bayes' rule to obtain the posterior:

$$P_G(T|I) = \frac{P(T, I)}{P(I)} \quad \text{(IV)}$$

$$\propto \sum_{\mathbf{w}} P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T) \quad \text{(V)}$$

Likewise, if we are focused on inferring which words were seen given an uncertain input, we have

$$P_G(\mathbf{w}|I) \propto \sum_T P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T) \quad \text{(VI)}$$

2.1 Uncertainty for a Known Input

This paper considers situations such as controlled psycholinguistic experiments where we (the researchers) know the sentence $\mathbf{w}^*$ presented to a comprehender, but do not know the specific input I that the comprehender obtains. In this case, if we are, for example, interested in the expected inferences of a rational comprehender about what word string she was exposed to, the probability distribution of interest is

$$P(\mathbf{w}|\mathbf{w}^*) = \int_I P_C(\mathbf{w}|I, \mathbf{w}^*)P_T(I|\mathbf{w}^*)\, dI \quad \text{(VII)}$$

where P_C is the probability distribution used by the comprehender to process perceived input, and P_T is the "true" probability distribution over the inputs

[1] By assumption, G is defined such that its complete productions T completely specify the string, such that $P(\mathbf{w}|T)$ is non-zero for only one value of $\mathbf{w}$.

that might actually be perceived given the true sentence. Since the comprehender does not observe $\mathbf{w}^*$ we must have conditional independence between $\mathbf{w}$ and $\mathbf{w}^*$ given I. We can then apply Bayes' rule to (VII) to obtain

$$P(\mathbf{w}|\mathbf{w}^*) = \int_I \frac{P_C(I|\mathbf{w})P_C(\mathbf{w})}{P_C(I)} P_T(I|\mathbf{w}^*)\, dI \quad \text{(VIII)}$$

$$= P_C(\mathbf{w}) \int_I \frac{P_C(I|\mathbf{w})P_T(I|\mathbf{w}^*)}{P_C(I)}\, dI \quad \text{(IX)}$$

$$\propto P_C(\mathbf{w})Q(\mathbf{w}, \mathbf{w}^*) \quad \text{(X)}$$

where $Q(\mathbf{w}, \mathbf{w}^*)$ is proportional to the integral term in Equation (IX). The term $P_C(\mathbf{w})$ corresponds to the comprehender's prior beliefs; the integral term is the effect of input uncertainty. If comprehenders model noise rationally, then we should have $P_C(I|\mathbf{w}) = P_T(I|\mathbf{w})$, and thus $Q(\mathbf{w}, \mathbf{w}^*)$ becomes a symmetric, non-negative function of $\mathbf{w}$ and $\mathbf{w}^*$; hence the effect of input uncertainty can be modeled by a *kernel function* on input string pairs. (Similar conclusions result when the posterior distribution of interest is over structures T.) It is an open question which kernel functions might best model the inferences made in human sentence comprehension. Most obviously the kernel function should account for noise (environmental, perceptual, and attentional) introduced into the signal en route to the neural stage of abstract sentence processing. In addition, this kernel function might also be a natural means of accounting for modeling error such as disfluencies (Johnson and Charniak, 2004), word/phrase swaps, and even well-formed utterances that the speaker did not intend. For purposes of this paper, we limit ourselves to a simple kernel based on the Levenshtein distance $LD(w, w')$ between words and constructed in the form of a weighted finite-state automaton (Mohri, 1997).

2.2 The Levenshtein-distance kernel

Suppose that the input word string $\mathbf{w}^*$ consists of words $w_{1\ldots n}$. We define the Levenshtein-distance kernel as follows. Start with a weighted finite-state automaton in the log semiring over the vocabulary Σ with states $0 \ldots n$, state 0 being the start state and n the (zero-cost) final state. We add two types of arcs to this automaton: (a) substitution/deletion arcs $(i-1, w') \rightarrow i$, $i \in 1, \ldots, n$, each with cost $\lambda\, LD(w_i, w')$, for all $w' \in \Sigma \cup \{\epsilon\}$; and (b) insertion loop arcs $(j, w') \rightarrow j$, $j \in 0, \ldots, n$, each with cost $\lambda\, LD(\epsilon, w')$, for all $w' \in \Sigma$.[2] The resulting wFSA $K_{LD}(\mathbf{w}^*)$ defines a function over $\mathbf{w}$ such that the summed weight of paths through the wFSA accepting $\mathbf{w}$ is $\log Q(\mathbf{w}, \mathbf{w}^*)$. This kernel allows for the possibility of word *substitutions* (represented by the transition arcs with labels that are neither w_i nor ϵ), word *deletions* (represented by the transition arcs with ϵ labels), and even word *insertions* (represented by the loop arcs). The unnormalized probability of each type of operation is exponential in the Levenshtein distance of the change induced by the operation. The term λ is a free parameter, with smaller values corresponding to noisier input. Figure 1 gives an example of the Levenshtein-distance kernel for a simple vocabulary and sentence.[3]

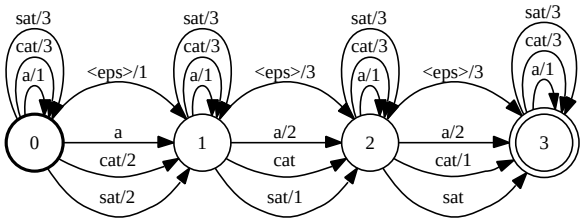

Figure 1: The Levenshtein-distance kernel for multi-word string edits. $K_{LD}(\mathbf{w}^*)$ is shown for $\Sigma = \{\text{cat,sat,a}\}$, $\mathbf{w}^* = (\text{a cat sat})$, and $\lambda = 1$. State 0 is the start state, and State 3 is the lone (zero-cost) final state.

[2] For purposes of computing the Levenshtein distance between words, the epsilon label ϵ is considered to be a zero-length letter string.

[3] The Levenshtein-distance kernel can be seen to be symmetric in $\mathbf{w}, \mathbf{w}^*$ as follows. Any path accepting $\mathbf{w}$ in the wFSA generated from $\mathbf{w}^*$ involves the following non-zero-cost transitions: insertions $w'^I_{1\ldots i}$, deletions $w^D_{1\ldots j}$, and substitutions $(w \rightarrow w')^S_{1\ldots k}$. For each such path P, there will be exactly one path P' in the wFSA generated from $\mathbf{w}$ that accepts $\mathbf{w}^*$ with insertions $w^D_{1\ldots j}$, deletions $w'^I_{1\ldots i}$, and substitutions $(w' \rightarrow w)^S_{1\ldots k}$. Due to the symmetry of the Levenshtein distance, the paths P and P' will have identical costs. Therefore the kernel is indeed symmetric.

2.3 Efficient computation of posterior beliefs

The problem of finding structures or strings with high posterior probability given a particular input string $\mathbf{w}^*$ is quite similar to the problem faced in the parsing of speech, where the acoustic input I to a parser can be represented as a lattice of possible word sequences, and the edges of the lattice have weights determined by a model of acoustic realization of words, $P(I|\mathbf{w})$ (Collins et al., 2004; Hall and Johnson, 2003, 2004). The two major differences between lattice parsing and our problem are (a) we have integrated out the expected effect of noise, which is thus implicit in our choice of kernel; and (b) the loops in the Levenshtein-distance kernel mean that the input to parsing is no longer a lattice. This latter difference means that some of the techniques applicable to string parsing and lattice parsing – notably the computation of inside probabilities – are no longer possible using exact methods. We return to this difference in Sections 3 and 4.

3 Global inference

One clear prediction of the uncertain-input model of (VII)–(X) is that under appropriate circumstances, the prior expectations $P_C(\mathbf{w})$ of the comprehender should in principle be able to override the linguistic input actually presented, so that a sentence is interpreted as meaning—and perhaps even *being*—something other than it actually meant or was. At one level, it is totally clear that comprehenders do this on a regular basis: the ability to do this is required for someone to act as a copy editor—that is, to notice and (crucially) correct mistakes on the printed page. In many cases, these types of correction happen at a level that may be below consciousness—thus we sometimes miss a typo but interpret the sentences as it was intended, or ignore the disfluency of a speaker. What has not been previously proposed in a formal model, however, is that this can happen *even when an input is a completely grammatical sentence*. Here, we argue that an effect demonstrated by Christianson et al. (2001) (see also Ferreira et al., 2002) is an example of expectations overriding input. When presented sentences of the forms in (1) using methods that did not permit rereading, and asked questions of the type *Did the man hunt the deer?*, experimental participants gave affirmative responses significantly more often for sentences of type (1a), in which the substring *the man hunted the deer* appears, than for either (1b) or (1c).

(1) a. While the man hunted the deer ran into the woods. (GARDENPATH)
b. While the man hunted the pheasant the deer ran into the woods. (TRANSITIVE)
c. While the man hunted, the deer ran into the woods. (COMMA)

This result was interpreted by Christianson et al. (2001) and Ferreira et al. (2002) as reflecting (i) the fact that there is a syntactic garden path in (1a)—after reading the first six words of the sentence, the preferred interpretation of the substring *the man hunted the deer* is as a simple clause indicating that the deer was hunted by the man—and (ii) that readers were not always successful at revising away this interpretation when they saw the disambiguating verb *ran*, which signals that *the deer* is actually the subject of the main clause, and that *hunted* must therefore be intransitive. Furthermore (and crucially), for (1a) participants also responded affirmatively most of the time to questions of the type *Did the deer run into the woods?* This result is a puzzle for existing models of sentence comprehension because no grammatical analysis exists of any substring of (1a) for which *the deer* is both the object of *hunted* and the subject of *ran*. In fact, no formal model has yet been proposed to account for this effect.

The uncertain-input model gives us a means of accounting for these results, because there are near neighbors of (1a) for which there *is* a global grammatical analysis in which either *the deer* or a coreferent NP is in fact the object of the subordinate-clause verb *hunted*. In particular, inserting the word *it* either before or after *the deer* creates such a near neighbor:

(2) a. While the man hunted the deer it ran into the woods.
b. While the man hunted it the deer ran into the woods.

We formalize this intuition within our model by using the wFSA representation of the Levenshtein-

ROOT	→ S PUNCT.	0.0
S	→ SBAR S	6.3
S	→ SBAR PUNCT_S	4.6
PUNCT_S	→ , S	0.0
S	→ NP VP	0.1
SBAR	→ IN S	0.0
NP	→ DT NN	1.9
NP	→ NNS	4.4
NP	→ NNP	3.3
NP	→ DT NNS	4.5
NP	→ PRP	1.3
NP	→ NN	3.1
VP	→ VBD RB	9.7
VP	→ VBD PP	2.2
VP	→ VBD NP	1.2
VP	→ VBD RP	8.3
VP	→ VBD	2.0
VP	→ VBD JJ	3.4
PP	→ IN NP	0.0

Figure 2: The PCFG used in the global-inference study of Section 3. Rule weights given as negative log-probabilities in bits.

distance kernel. A probabilistic context-free grammar (PCFG) representing the comprehender's grammatical knowledge can be intersected with that wFSA using well-understood techniques, generating a new weighted CFG (Bar-Hillel et al., 1964; Nederhof and Satta, 2003). This intersection thus represents the unnormalized posterior $P_C(T, \mathbf{w}|\mathbf{w}^*)$. Because there are loops in the wFSA generated by the Levenshtein-distance kernel, exact normalization of the posterior is not tractable (though see Nederhof and Satta, 2003; Chi, 1999; Smith and Johnson, 2007 for possible approaches to approximating the normalization constant). We can, however, use the lazy k-best algorithm of Huang and Chiang (2005; Algorithm 3) to obtain the word-string/parse-tree pairs with highest posterior probability.

3.1 Experimental Verification

To test our account of the rational noisy-channel interpretation of sentences such as (1), we defined a small PCFG using the phrasal rules listed in Figure 2, with rule probabilities estimated from the parsed Brown corpus.[4] Lexical rewrite probabilities were determined using relative-frequency estimation over the entire parsed Brown corpus. For each of the sentence sets like (1) used in Experiments 1a, 1b, and 2 of Christianson et al. (2001) that have complete lexical coverage in the parsed Brown corpus (22 sets in total), a noisy-input wFSA was constructed using K_{LD}, permitting all words occurring more than 2500 times in the parsed Brown corpus as possible edit/insertion targets.[5] Figure 3 shows the average proportion of parse trees among the 100 best parses in the intersection between this PCFG and the wFSA for each sentence for which an interpretation is available such that *the deer* or a coreferent NP is the direct object of *hunted*.[6] The Levenshtein distance penalty λ is a free parameter in the model, but the results are consistent for a wide range of λ: interpretations of type (2) are more prevalent both in terms of number mass for (1a) than for either (1b) or (1c). Furthermore, across 9 noise values for 22 sentence sets, there were never more interpretations of type (2) for COMMA sentences than for the corresponding GARDENPATH sentences, and in only one case were there more such interpretations for a TRANSITIVE sentence than for the corresponding GARDENPATH sentence.

4 Incremental comprehension and error identification

We begin taking up the role of input uncertainty for incremental comprehension by posing a question:

[4] Counts of these rules were obtained using `tgrep2/Tregex` tree-matching patterns (Rohde, 2005; Levy and Andrew, 2006), available online at `http://idiom.ucsd.edu/~rlevy/papers/emnlp2008/tregex_patterns`. We have also investigated the use of broad-coverage PCFGs estimated using standard treebank-based techniques, but found that the computational cost of inference with treebank-sized grammars was prohibitive.

[5] The word-frequency cutoff was introduced for computational speed; we have obtained qualitatively similar results with lower word-frequency cutoffs.

[6] We took a parse tree to satisfy this criterion if the NP *the deer* appeared either as the matrix-clause subject or the embedded-clause object, and a pronoun appeared in the other position. In a finer-grained grammatical model, number/gender agreement would be enforced between such a pronoun and the NP in the posterior, so that the probability mass for these parses would be concentrated on cases where the pronoun is *it*.

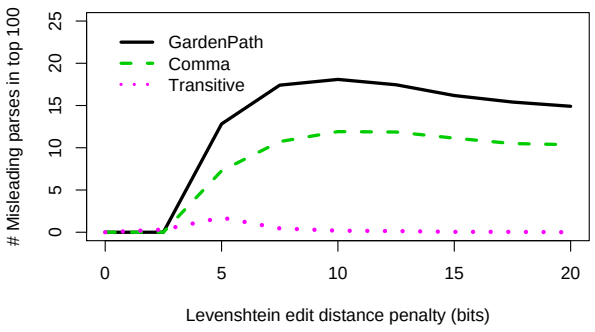

Figure 3: Results for 100-best global inference, as a function of the Levenshtein distance penalty λ (in bits).

what is the optimal way to read a sentence on a page (Legge et al., 1997)? Presumably, the goal of reading is to find a good compromise between scanning the contents of the sentence as quickly as possible while achieving an accurate understanding of the sentence's meaning. To a first approximation, humans solve this problem by reading each sentence in a document from beginning to end, regardless of the actual layout; whether this general solution is best understood in terms of optimality or rather as parasitic on spoken language comprehension is an open question beyond the immediate scope of the present paper. However, about 10–15% of eye movements in reading are regressive (Rayner, 1998), and we may usefully refine our question to when a *regressive* eye movement might be a good decision. In traditional models of sentence comprehension, the optimal answer would certainly be "never", since past observations are known with certainty. But once uncertainty about the past is accounted for, it is clear that there may in principle be situations in which regressive saccades may be the best choice.

What are these situations? One possible answer would be: when the uncertainty (e.g., measured by entropy) about an earlier part of the sentence is high. There are some cases in which this is probably the correct answer: many regressive eye movements are very small and the consensus in the eye-movement literature is that they represent corrections for motor error at the saccadic level. That is, the eyes overshoot the intended target and regress to obtain information about what was missed. However, motor error can account only for short, isolated regressions, and about one-sixth of regressions are part of a longer series back into the sentence, into a much earlier part of the text which has already been read. We propose that these regressive saccades might be the best choice *when the most recent observed input significantly changes the comprehender's beliefs about the earlier parts of the sentence*. To make the discussion more concrete, we turn to another recent result in the psycholinguistic literature that has been argued to be problematic for rational theories of sentence comprehension.

It has been shown (Tabor et al., 2004) that sentences such as (3) below induce considerable processing difficulty at the word *tossed*, as measured in word-by-word reading times:

(3) The coach smiled at the player tossed a frisbee. (LOCALLY COHERENT)

Both intuition and controlled experiments reveal that this difficulty seems due at least in part to the category ambiguity of the word *tossed*, which is occasionally used as a participial verb but is much more frequently used as a simple-past verb. Although *tossed* in (3) is actually a participial verb introducing a reduced relative clause (and *the player* is hence its recipient), most native English speakers find it extremely difficult not to interpret *tossed* as a main verb and *the player* as its agent—far more difficult than for corresponding sentences in which the critical participial verb is morphologically distinct from the simple past form ((4a), (4c); c.f. *threw*) or in which the relative clause is unreduced and thus clearly marked ((4b), (4c)).

(4) a. The coach smiled at the player thrown a frisbee. (LOCALLY INCOHERENT)
b. The coach smiled at the player who was tossed a frisbee.
c. The coach smiled at the player who was thrown a frisbee.

The puzzle here for rational approaches to sentence comprehension is that the preceding top-down context provided by *The coach smiled at...* should completely rule out the possibility of seeing a main verb immediately after *player*, hence a rational com-

prehender should not be distracted by the part-of-speech ambiguity.[7]

4.1 An uncertain-input solution

The solution we pursue to this puzzle lies in the fact that (3) has many near-neighbor sentences in which the word *tossed* is in fact a simple-past tense verb. Several possibilities are listed below in (5):

(5)
a. The coach **who** smiled at the player tossed a frisbee.
b. The coach smiled **as** the player tossed a frisbee.
c. The coach smiled **and** the player tossed a frisbee.
d. The coach smiled at the player **who** tossed a frisbee.
e. The coach smiled at the player **that** tossed a frisbee.
f. The coach smiled at the player **and** tossed a frisbee.

The basic intuition we follow is that simple-past verb *tossed* is much more probable where it appears in any of (5a)-(5f) than participial *tossed* is in (3). Therefore, seeing this word causes the comprehender to shift her probability distribution about the earlier part of the sentence away from (3), where it had been peaked, toward its near neighbors such as the examples in (5). This change in beliefs about the past is treated as an error identification signal (EIS). In reading, a sensible response to an EIS would be a slowdown or a regressive saccade; in spoken language comprehension, a sensible response would be to allocate more working memory resources to the comprehension task.

4.2 Quantifying the Error Identification Signal

We quantify our proposed error identification signal as follows. Consider the probability distribution over the input up to, but not including, a position j in a sentence $\mathbf{w}$:

$$P(w_{[0,j)}) \tag{XI}$$

We use the subscripting $[0, j)$ to illustrate that this interval is "closed" through to include the beginning of the string, but "open" at position j—that is, it includes all material before position j but does not include anything at that position or beyond. Let us then define the posterior distribution after seeing all input up through and including word i as $P_i(w_{[0,j)})$. We define the EIS induced by reading a word w_i as follows:

$$D\left(P_i(w_{[0,i)})||P_{i-1}(w_{[0,i)})\right) \tag{XII}$$

$$\equiv \sum_{w \in \left\{w_{[0,i)}\right\}} P_i\left(w\right) \log \frac{P_i\left(w\right)}{P_{i-1}\left(w\right)} \tag{XIII}$$

where $D(q||p)$ is the Kullback-Leibler divergence, or relative entropy, from p to q, a natural way of quantifying the distance between probability distributions (Cover and Thomas, 1991) which has also been argued for previously in modeling attention and surprise in both visual and linguistic cognition (Itti and Baldi, 2005; Levy, 2008).

4.3 Experimental Verification

As in Section 3, we use a small probabilistic grammar covering the relevant structures in the problem domain to represent the comprehender's knowledge, and a wFSA based on the Levenshtein-distance kernel to represent noisy input. We are interested in comparing the EIS at the word *tossed* in (3) versus the EIS at the word *thrown* in (4a). In this case, the interval $w_{[0,j)}$ contains all the material that could possibly have come before the word *tossed/thrown*, but does not contain material at or after the position introduced by the word itself. Loops in the probabilistic grammar and the Levenshtein-distance kernel pose a challenge, however, to evaluating the EIS, because the normalization constant of the resulting grammar/input intersection is essential to evaluating Equation (XIII). To circumvent this problem, we eliminate loops from the kernel by allowing only one insertion per inter-word space.[8] (See Section 5 for a possible alternative).

[7] This preceding context sharply distinguishes (3) from better-known, traditional garden-path sentences such as *The horse raced past the barn fell*, in which preceding context cannot be used to correctly disambiguate the part of speech of the ambiguous verb *raced*.

[8] Technically, this involves the following transformation of a Levenshtein-distance wFSA. First, eliminate all loop arcs.

ROOT	→ S	0.00
S	→ S-base CC S-base	7.3
S	→ S-base	0.01
S-base	→ NP-base VP	0
NP	→ NP-base RC	4.1
NP	→ NP-base	0.5
NP	→ NP-base PP	2.0
NP-base	→ DT N N	4.7
NP-base	→ DT N	1.9
NP-base	→ DT JJ N	3.8
NP-base	→ PRP	1.0
NP-base	→ NNP	3.1
VP/NP	→ V NP	4.0
VP/NP	→ V	0.1
VP	→ V PP	2.0
VP	→ V NP	0.7
VP	→ V	2.9
RC	→ WP S/NP	0.5
RC	→ VP-pass/NP	2.0
RC	→ WP FinCop VP-pass/NP	4.9
PP	→ IN NP	0
S/NP	→ VP	0.7
S/NP	→ NP-base VP/NP	1.3
VP-pass/NP	→ VBN NP	2.2
VP-pass/NP	→ VBN	0.4

Figure 4: The grammar used for the incremental-inference experiment of Section 4. Rule weights given as negative log-probabilities in bits.

Figure 4 shows the (finite-state) probabilistic grammar used for the study, with rule probabilities once again determined from the parsed Brown corpus using relative frequency estimation. To calculate the distribution over strings after exposure to the i-th word in the sentence, we "cut" the input wFSA such that all transitions and arcs after state $2i+2$ were removed and replaced with a sequence of states $j = 2i+3, \ldots, m$, with zero-cost transitions $(j-1, w') \rightarrow j$ for all $w' \in \Sigma \cup \{\epsilon\}$, and each new j being a zero-cost final state.[9] Because the intersection between this "cut" wFSA and the probabilistic grammar is loop-free, it can be renormalized, and the EIS can be calculated without difficulty. All the computations in this section were carried out using the OpenFST library (Allauzen et al., 2007).

Next, map every state i onto a state pair in a new wFSA $(2i, 2i+1)$, with all incoming arcs in i being incoming into $2i$, all outgoing arcs from i being outgoing from $2i+1$, and new transition arcs $(2i, w') \rightarrow 2i+1$ for each $w' \in \Sigma \cup \{\epsilon\}$ with cost $LD(\epsilon, w')$. Finally, add initial state 0 to the new wFSA with transition arcs to state 1 for all $w' \in \Sigma \cup \{\epsilon\}$ with cost $LD(\epsilon, w')$. A final state i in the old wFSA corresponds to a final state $2i+1$ in the new wFSA.

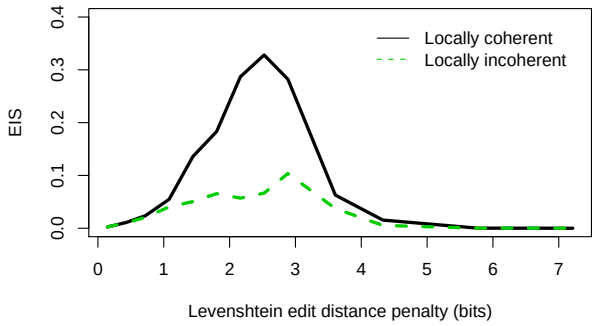

Figure 5: The Error Identification Signal (EIS) for (3) and (4a), as a function of the Levenshtein distance penalty λ (in bits)

Figure 5 shows the average magnitude of the EIS for sentences (3) versus (4a) at the critical word position *tossed/thrown*. Once again, the Levenshtein-distance penalty λ is a free parameter in the model, so we show model behavior as a function of λ, for the eight sentence pairs in Experiment 1 of Tabor et al. with complete lexical and syntactic coverage for the grammar of Figure 4. For values of λ where the EIS is non-negligible, it is consistently larger at the critical word (*tossed* in (3), *thrown* in (4a)) in the COHERENT condition than in the INCOHERENT condition. Across a range of eight noise levels, 67% of sentence pairs had a higher EIS in the COHERENT condition than in the INCOHERENT condition. Furthermore, the cases where the INCOHERENT condition had a larger EIS occurred only for noise levels below 1.1 and above 3.6, and the maximum such EIS was quite small, at 0.067. Overall, the model's behavior is consistent with the experimental results of Tabor et al. (2004), and can be explained through the intuition described at the end of Section 4.1.

[9] The number of states added had little effect on results, so long as at least as many states were added as words remained in the sentence.

5 Conclusion

In this paper we have outlined a simple model of rational sentence comprehension under uncertain input and explored some of the consequences for outstanding problems in the psycholinguistic literature. The model proposed here will require further empirical investigation in order to distinguish it from other proposals that have been made in the literature, but if our proposal turns out to be correct it has important consequences for both the theory of language processing and cognition more generally. Most notably, it furthers the case for rationality in sentence processing; and it eliminates one of the longest-standing modularity hypotheses implicit in work on the cognitive science of language: a partition between systems of word recognition and sentence comprehension (Fodor, 1983). Unlike the pessimistic picture originally painted by Fodor, however, the interactivist picture resulting from our model's joint inference over possible word strings and structures points to many rich details that still need to be filled in. These include questions such as what kernel functions best account for human comprehenders' modeling of noise in linguistic input, and what kinds of algorithms might allow representations with uncertain input to be computed incrementally.

The present work could also be extended in several more technical directions. Perhaps most notable is the problem of the normalization constant for the posterior distribution over word strings and structures; this problem was circumvented via a k-best approach in Section 3 and by removing loops from the Levenshtein-distance kernel in Section 4. We believe, however, that a more satisfactory solution may exist via sampling from the posterior distribution over trees and strings. This may be possible either by estimating normalizing constants for the posterior grammar using iterative weight propagation and using them to obtain proper production rule probabilities (Chi, 1999; Smith and Johnson, 2007), or by using reversible-jump Markov-chain Monte Carlo (MCMC) techniques to sample from the posterior (Green, 1995), and estimating the normalizing constant with annealing-based techniques (Gelman and Meng, 1998) or nested sampling (Skilling, 2004). Scaling the model up for use with treebank-size grammars is another area for technical improvement.

Finally, we note that the model here could potentially find practical application in grammar correction. Although the noisy channel has been in use for many years in spelling correction, our model could be used more generally for grammar corrections, including insertions, deletions, and (with new noise functions) potentially changes in word order.

References

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*, volume 4783 of *Lecture Notes in Computer Science*, pages 11–23. Springer. `http://www.openfst.org`.

Altmann, G. T. and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Aylett, M. and Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Bar-Hillel, Y., Perles, M., and Shamir, E. (1964). On formal properties of simple phrase structure grammars. In *Language and Information: Selected Essays on their Theory and Application*. Addison-Wesley.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of AAAI*, pages 598–603.

Chi, Z. (1999). Statistical properties of probabilistic context-free grammars. *Computational Linguistics*, 25(1):131–160.

Christianson, K., Hollingworth, A., Halliwell, J. F., and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42:368–407.

Collins, C., Carpenter, B., and Penn, G. (2004). Head-driven parsing for word lattices. In *Proceedings of ACL*.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley.

Ferreira, F., Ferraro, V., and Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11:11–15.

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

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.

Genzel, D. and Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of ACL*.

Genzel, D. and Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In *Empirical Methods in Natural Language Processing*, volume 10.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo and Bayesian model determination. *Biometrika*, 82:711–732.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, volume 2, pages 159–166.

Hall, K. and Johnson, M. (2003). Language modeling using efficient best-first bottom-up parsing. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*.

Hall, K. and Johnson, M. (2004). Attention shifting for parsing speech. In *Proceedings of ACL*.

Huang, L. and Chiang, D. (2005). Better k-best parsing. In *Proceedings of the International Workshop on Parsing Technologies*.

Itti, L. and Baldi, P. (2005). Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems*.

Johnson, M. and Charniak, E. (2004). A TAG-based noisy channel model of speech repairs. In *Proceedings of ACL*.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2):137–194.

Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 317–324, Barcelona.

Legge, G. E., Klitz, T. S., and Tjan, B. S. (1997). Mr. Chips: An ideal-observer model of reading. *Psychological Review*, 104(3):524–553.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177.

Levy, R. and Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the 2006 conference on Language Resources and Evaluation*.

Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2):269–311.

Narayanan, S. and Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading time in sentence processing. In *Advances in Neural Information Processing Systems*, volume 14, pages 59–65.

Nederhof, M.-J. and Satta, G. (2003). Probabilistic parsing as intersection. In *Proceedings of the International Workshop on Parsing Technologies*.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Rohde, D. (2005). *TGrep2 User Manual*, version 1.15 edition.

Skilling, J. (2004). Nested sampling. In Fischer, R., Preuss, R., and von Toussaint, U., editors, *Bayesian inference and maximum entropy methods in science and engineering*, number 735 in AIP Conference Proceedings, pages 395–405.

Smith, N. A. and Johnson, M. (2007). Weighted and probabilistic context-free grammars are equally expressive. *Computational Linguistics*, 33(4):477–491.

Tabor, W., Galantucci, B., and Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4):355–370.

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