A model of local coherence effects in human sentence processing as consequences of updates from bottom-up prior to posterior beliefs

Klinton Bicknell and **Roger Levy**
Department of Linguistics
University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093-0108
{kbicknell,rlevy}@ling.ucsd.edu

Abstract

Human sentence processing involves integrating probabilistic knowledge from a variety of sources in order to incrementally determine the hierarchical structure for the serial input stream. While a large number of sentence processing effects have been explained in terms of comprehenders' rational use of probabilistic information, effects of local coherences have not. We present here a new model of local coherences, viewing them as resulting from a belief-update process, and show that the relevant probabilities in our model are calculable from a probabilistic Earley parser. Finally, we demonstrate empirically that an implemented version of the model makes the correct predictions for the materials from the original experiment demonstrating local coherence effects.

1 Introduction

The task of human sentence processing, recovering a hierarchical structure from a serial input fraught with local ambiguities, is a complex and difficult problem. There is ample evidence that comprehenders understand sentences incrementally, constructing interpretations of partial structure and expectations for future input (Tanenhaus et al., 1995; Altmann and Kamide, 1999). Many of the main behavioral findings in the study of human sentence processing have now been explained computationally. Using probabilistic models trained on large-scale corpora, effects such as global and incremental disambiguation preferences have been shown to be a result of the rational use of syntactic probabilities (Jurafsky, 1996; Hale, 2001; Narayanan and Jurafsky, 2001; Levy, 2008b; Levy et al., 2009). Similarly, a number of other effects in both comprehension and production have been modeled as resulting from rational strategies of languages users that take into account all the probabilistic information present in the linguistic signal (Genzel and Charniak, 2002; Genzel and Charniak, 2003; Keller, 2004; Levy and Jaeger, 2007).

One class of results from the literature that has not yet been explained in terms of a rational comprehender strategy is that of local coherence effects (Tabor et al., 2004; Gibson, 2006; Konieczny and Müller, 2007), cases in which it appears that the parser is systematically ignoring contextual information about possible syntactic structures and pursuing analyses that are probable only locally. These effects are problematic for rational models, because of the apparent failure to use all of the available information. This paper describes a new model of local coherence effects under rational syntactic comprehension, which proposes that they arise as a result of updating prior beliefs about the structures that a given string of words is likely to have to posterior beliefs about the likelihoods of those structures in context. The critical intuition embodied in the model is that larger updates in probability distributions should be more processing-intensive; hence, the farther the posterior is from the prior, the more radical the update required and the greater the processing load. Section 2 describes the problem of local coherences in detail and Section 3 describes existing models of the phenomenon. Following that, Sections 4–5 describe our model and its computa-

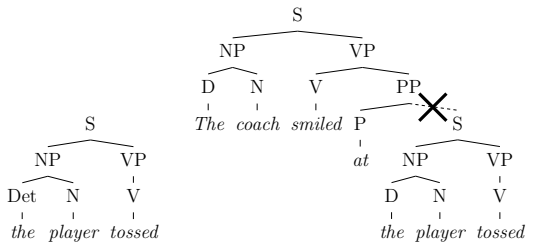

Figure 1: The difficulty of explaining local-coherence effects as traditional garden-pathing.

tion from a probabilistic Earley parser. Section 6 presents the results of an experiment showing that our model makes the correct predictions for the local coherence effects seen in the original paper by Tabor et al. (2004). Finally, Section 7 concludes and discusses the insight our model gives into human performance.

2 Local coherences

The first studies to report effects of local coherences are described in Tabor et al. (2004). In Experiment 1, they use a self-paced reading task and materials containing relative clauses (RCs) attached to nouns in non-subject position as in (1).

(1) a. The coach smiled at the player tossed a frisbee by the opposing team.
b. The coach smiled at the player who was tossed a frisbee by the opposing team.
c. The coach smiled at the player thrown a frisbee by the opposing team.
d. The coach smiled at the player who was thrown a frisbee by the opposing team.

Their experimental design crossed RC reduction with verb ambiguity. RCs are either reduced (1a,1c) or unreduced (1b,1d), and the RC verb is either lexically ambiguous between a past tense active and a past participle (1a–1b), or is unambiguously a past participle (1c–1d).

Tabor et al. point out that in one of these four conditions (1a) there is a locally coherent string *the player tossed a frisbee*. Out of context (e.g., if it were starting a sentence) this string would have a likely parse in which *tossed* is a past tense active verb, *the player* is its agent, and *a frisbee* is its theme (Figure 1, left). The preceding context within the sentence, however, should rule out this interpretation because *the player* appears within a PP and hence should not be able to be the subject of a new sentence (Figure 1, right). That is, given the preceding context, *the player tossed a frisbee* must begin a reduced RC, such that there is no local ambiguity. Thus, if comprehenders are making full use of the linguistic context, (1a) should be no more difficult than the other examples, except insofar as ambiguous verbs are harder than unambiguous verbs, and reduced RCs are harder than unreduced RCs, predicting there would be only the two main effects of RC reduction and verb ambiguity on reading times for the *tossed a frisbee* region.

Tabor et al., however, predict an interaction such that (1a) will have added difficulty above and beyond these two effects, because of the interference from the locally coherent parse of *the player tossed a frisbee*. Concordant with their predictions, they find an interaction in the *tossed a frisbee* region, such that (1a) is super-additively difficult. Because this result requires that an impossible parse influences a word's difficulty, it is in direct opposition to the predictions of theories of processing difficulty in which the probability of a word given context is the primary source of parsing difficulty, and more generally appears to be in opposition to any rational theory, in which comprehenders are making use of all the information in the linguistic context.

3 Existing models

With the results showing local coherence effects in mind, we can ask the question of what sorts of theories do predict these effects. This section briefly describes two recent examples of such theories. The first involves dynamical systems models to explain the effects and the second uses a mathematical model of the combination of bottom-up and top-down probabilistic information.

In Tabor and Hutchins's (2004) SOPARSE (self-organized parse) model, reading a word activates a set of lexically anchored tree fragments. Through spreading activation between compatible fragments and inhibition between incompatible ones, these tree fragments then compete in a process which is sensitive only to the local environment, i.e., ignoring the global grammatical context. Eventually, the sys-

tem stabilizes to the correct parse, and reading times for each word are modeled as the time the system takes to stabilize after reading a word. Stabilization takes longer for locally coherent regions because the locally coherent parse will be created and compete with the globally grammatical parse.

There are, however, unresolved issues with this model. The model has a number of free parameters, relating to the equations used for the competition, the method by which links between fragments are formed, as well as the question of precisely what tree fragments a given word will activate. While Tabor and Hutchins (2004) work out these questions in detail for the types of sentences they model, it is unclear how the model could be scaled up to make predictions for arbitrary types of sentences. That is, there is no principled system for setting the three types of parameters mentioned, and thus no clear interpretation of their values. The model put forward in this paper is an attempt to remedy this situation.

A recent proposal by Gibson (2006) can also explain some of the local coherence results. Gibson's proposal is that part-of-speech ambiguities have a special status in parsing; in effect, lexical part-of-speech ambiguities can be thought of as one-word local coherences. In this model, a probability function $\tilde{P}$ is calculated over part-of-speech tags given a word. This probability for tag t_i and a word w, $\tilde{P}(t_i|w)$, is proportional to the context-independent probability of t_i given the word w, $P(t_i|w)$ – the *bottom-up* component – multiplied by a smoothed probability P_s of the tag given the context – the *top-down* component:

$$\tilde{P}(t_i|w) = \frac{P(t_i|w)P_s(t_i|\text{context})}{\sum_{t \in T} P(t|w)P_s(t|\text{context})} \qquad (1)$$

Difficulty is predicted to be high when the probability $\tilde{P}$ of the correct tag is low.

Because the top-down probabilities are smoothed to allow for all possible parts-of-speech, any word which is lexically ambiguous will be more difficult to process, regardless of whether it is ambiguous or not in its context. This can thus explain some of the difference between the ambiguous and unambiguous verbs in Tabor et al. (2004). It is not clear, however, under such a model why the super-additive interaction would obtain—that is, why (1a) should be so much harder than (1b) starting at the word *tossed*. In addition, Gibson's model is a bit underspecified: he does not discuss how the top-down probabilities are calculated, nor what the precise linking hypothesis is between the final $\tilde{P}$ and reading times. Finally, it is not at all clear why the top-down expectations should be smoothed, since the smoothing actually has negative consequences on the processor's performance.

4 Parsing as belief update

The basic intuition behind the model presented here is that incrementally processing a sentence can be conceptualized as a process of updating one's beliefs. Such an analogy has been used to motivate surprisal-based theories of sentence processing (Hale, 2001; Levy, 2008a), where beliefs about the structure of a sentence after seeing the first $i-1$ words in the sentence, which we denote as w_0^{i-1}, are updated upon encountering w_i. In this case, the *surprisal* of a word ($-\log P(w_i|w_0^{i-1})$) is equivalent to the Kullback-Leibler divergence of the beliefs after w_0^i from the beliefs after w_0^{i-1} (Levy, 2008a). Our model focuses on another belief-update process in sentence processing: updating beliefs about the structures that a string of words is likely to have independent of context to beliefs about what structures it is likely to have in context. A bit more formally, it views the process of integrating a string of words w_i^j into a sentence as beginning with a 'bottom-up' prior distribution of syntactic structures likely to span w_i^j and integrating that with 'top-down' knowledge from the previous words in the sentence w_0^i in order to reach a posterior distribution conditioning on w_0^j over which structures actually can span w_i^j. This belief update process can be viewed as a rational reconstruction of the Tabor and Hutchins (2004) model, where – instead of the system dynamics of competition between arbitrary tree fragments – differences between prior and posterior probability distributions over syntactic structures determine processing difficulty.

More formally still, when integrating w_i^j into a sentence, for each syntactic category X, we can define the prior probability conditioned only on w_i^j that w_i^j will form the beginning of that category, i.e., that an X exists which begins at index i and spans at least

through j:

$$\text{Prior:} \quad P(X_i^{k \geq j} | w_i^j) \tag{2}$$

It is important to note here that this prior probability is conditional only on the value of w_i^j and not the values of i or j; that is, in the prior probability, i and j should be interpreted merely as a way to coindex the start and end points of the string of words being integrated with a category X potentially spanning them, and not as making reference to position in the full sentence string.

For each category X, this prior probability will be updated to the posterior probability of that category spanning w_i^j given all the words seen so far:

$$\text{Posterior:} \quad P(X_i^{k \geq j} | w_0^j) \tag{3}$$

In the equation for the posterior, of course, the indices i and j are positions in the sentence string, and not merely coindices.

Given these prior and posterior beliefs, we predict difficulty to arise in cases where the prior requires substantial modification to reach the posterior, that is, cases in which the prior and posterior make substantially different predictions for categories. A strong local coherence will have sharply different prior and posterior distributions, causing difficulty. We represent the prior and posterior beliefs as vectors of the probabilities of each syntactic category spanning w_i^j, and measure M_{ij}, the amount of modification required, as the summed K-L divergence of the prior from the posterior vector. That is, if N is the set of nonterminals in the grammar, the size of the belief update is modeled as[1]

$$M_{ij} \stackrel{\text{def}}{=} \sum_{X \in N} D\left(P(X_i^{k \geq j} | w_0^j) \,||\, P(X_i^{k \geq j} | w_i^j)\right)$$

In the remainder of the paper, we show how to compute M_{ij} by using Bayesian inference on quantities calculated in ordinary probabilistic incremental Earley parsing with a stochastic context-free grammar (SCFG), and show that our model makes the correct predictions using an SCFG for English on the original local-coherences experiment of Tabor et al. (2004).

[1]Note that for each syntactic category $X \in N$, the probability distribution $P(X_i^{k \geq j} | I)$ for some information I is over a binary random variable indicating the presence of X. The different syntactic categories X that could span from i to any k are not mutually exclusive, hence we cannot define size of belief update as a single K-L divergence defined over multinomial distributions.

5 Computing priors and posteriors

For SCFGs, a probabilistic Earley parser (Earley, 1970; Stolcke, 1995) provides the basic quantities we need to compute the prior (2) and posterior (3) for each category X. Following Stolcke, we use capital Latin characters to denote non-terminal categories and use lowercase Greek characters to denote (possibly null) sequences of terminals and non-terminals. We write the probability that a non-terminal X can be recursively rewritten by SCFG rules as a certain series of symbols μ by

$$P(X \stackrel{*}{\Rightarrow} \mu)$$

An edge built from the rule $X \rightarrow \lambda\mu$ where λ has been recognized as beginning at position i and ending at position j is denoted

$$j : X_i \rightarrow \lambda . \mu$$

The forward probability of that edge at position j, α_j, is defined to be the joint probability that the root node will generate all words recognized so far w_0^j as well as the edge

$$\alpha_j(X_i \rightarrow \lambda . \mu)$$

With this terminology, we are now in a position to describe how we calculate the posterior and prior probability vectors for our model.

5.1 Calculating the posterior

To calculate the posterior, we first use the definition of conditional probability to rewrite it as

$$P(X_i^{k \geq j} | w_0^j) = \frac{P(X_i^{k \geq j}, w_0^j)}{P(w_0^j)}$$

In a context-free grammar, given the syntactic category that dominates a string of words, the words' probability is independent from everything outside the category. Thus, this is equivalent to

$$\begin{aligned} P(X_i^{k \geq j} | w_0^j) &= \frac{P(w_o^i, X_i) P(w_i^j | X_i^{k \geq j})}{P(w_0^j)} \\ &= \frac{P(S \stackrel{*}{\Rightarrow} w_0^i X \nu) P(X \stackrel{*}{\Rightarrow} w_i^j \mu)}{P(S \stackrel{*}{\Rightarrow} w_0^j \lambda)} \end{aligned}$$

5.1.1 Posterior: the numerator's first term

The first term in the numerator $P(S \stackrel{*}{\Rightarrow} w_0^i X \nu)$ can be computed from a parse of w_0^i by summing forward probabilities of the form

$$\alpha_i(X_i \rightarrow .\mu) \tag{4}$$

5.1.2 Posterior: the denominator

Similarly, the denominator $P(S \stackrel{*}{\Rightarrow} w_0^j \lambda)$ can be computed from a parse of w_0^j by summing forward probabilities of the form

$$\alpha_j(Y \rightarrow \lambda w_{j-1}^j .\mu) \tag{5}$$

for all Y. This is because the forward probability of a state is conditioned on generating all the previous words.

5.1.3 Posterior: the numerator's second term

The second term in the numerator $P(X \stackrel{*}{\Rightarrow} w_i^j \mu)$ for an arbitrary category X cannot necessarily be calculated from a probabilistic Earley parse of the sentence, because the parser does not construct states that are not potentially useful in forming a sentence (i.e., states that would have a forward probability of zero.) However, to calculate the probability of X generating words w_i^j we can parse w_i^j separately with a goal category of X. From this parse, we can extract the probability of w_i^j being generated from X in the same way as we extracted the probability of w_0^j being generated from S, i.e., as a sum of forward probabilities at j (Eq. 5).[2]

5.2 Calculating the prior

To calculate the prior, we first use Bayes rule to rewrite it as

$$P(X_i^{k \geq j} | w_i^j) = \frac{P(w_i^j | X_i^{k \geq j}) P(X_i^{k \geq j})}{P(w_i^j)} \tag{6}$$

Recall that at this point, i and j do not refer to index positions in the actual string but rather serve to identify the substring w_i^j of interest. That is, $P(w_i^j)$ denotes the probability that at an arbitrary point in an arbitrary sentence, the next $j - i$ words will be w_i^j, and $P(X_i^{k \geq j})$ denotes the probability that an arbitrary point in an arbitrary sentence will be the left edge of a category X that spans at least $j - i$ words. None of the three terms in Eq. 6 can be directly obtained. However, we can obtain a very good approximation of Eq. 6 as follows. First, we marginalize over the position within a sentence with which the left edge i might be identified:

$$P(X_i^{k \geq j} | w_i^j) = \sum_{i'=0,1,\ldots} \left(\frac{P(w_{i'}^j | X_{i'}^{k \geq j}) P(X_{i'}^{k \geq j})}{P(w_{i'}^j)} \right) P(i = i') \tag{7}$$

In Eq. 7, i' *is* identified with the actual string position within the sentence.

Second, we rewrite the first term in this sum with event space notation, using the event space given in Table 1.

Table 1: Event space for the prior

Event	Description	
E_0:	There are at least i' words	$\lvert w \rvert \geq i'$
E_1:	A category X begins at i'	$X_{i'}$
E_2:	An $X_{i'}$ spans at least through j	$X_{i'}^{k \geq j}$
E_3:	There are at least j words	$\lvert w \rvert \geq j$
E_4:	Words $w_{i'}^j$ are these specific $\tilde{w}_{i'}^j$	$w_{i'}^j = \tilde{w}_{i'}^j$

$$\frac{P(w_{i'}^j | X_{i'}^{k \geq j}) P(X_{i'}^{k \geq j})}{P(w_{i'}^j)} = \frac{P(E_{0,3,4} | E_{0\ldots3}) P(E_{0\ldots3})}{P(E_{0,3,4})} = \frac{P(E_4 | E_{0\ldots3}) P(E_{o\ldots3})}{P(E_{o,3,4})}$$

Applying the chain rule, we can further simplify.

$$= \frac{P(E_4 | E_{0\ldots3}) P(E_{1\ldots3} | E_0) P(E_0)}{P(E_{3,4} | E_0) P(E_0)} = \frac{P(E_4 | E_{0\ldots3}) P(E_{1\ldots3} | E_0)}{P(E_{3,4} | E_0)} = \frac{P(E_{2\ldots4} | E_0, E_1) P(E_1 | E_0)}{P(E_{3,4} | E_0)}$$

[2] To calculate the posterior, it is not necessary to parse w_i^j separately, since these states are only excluded from the parse when their forward probability is zero, in which case the first term in the numerator will also be zero. A separate parse is necessary, however, when using this term to calculate the prior.

Switching back from event space notation and substituting this term into Eq. 7, we now have

$$P(X_i^{k\geq j}|w_i^j) = \sum_{i'=0,1,\ldots} \left(\frac{P(w_{i'}^j|X_{i'},E_0)P(X_{i'}|E_0)}{P(w_{i'}^j|E_0)} \right) P(i=i') \quad (8)$$

Thus, by conditioning all terms on E_0, the presence of at least i' words, we have transformed the probabilities we need to calculate into these four terms, which are easier to calculate from the parser. We now consider how to calculate each of the terms.

5.2.1 Prior: the numerator's first term

The first term in the numerator can be simplified because our grammar is context-free:

$$\begin{aligned} P(w_{i'}^j|X_{i'},E_0) &= P(w_{i'}^j|X_{i'}) \\ &= P(X \stackrel{*}{\Rightarrow} w_{i'}^j) \end{aligned}$$

This can be computed as described in Section 5.1.3.

5.2.2 Prior: the numerator's second term

The second term in the numerator can be rewritten as follows:

$$\begin{aligned} P(X_{i'}|E_0) &= \frac{P(X_{i'},E_0)}{P(E_0)} \\ &= \frac{P(S \stackrel{*}{\Rightarrow} \mathring{w}_0^{i'} X \mu)}{P(S \stackrel{*}{\Rightarrow} \mathring{w}_0^{i'} \mu)} \end{aligned}$$

where $\mathring{w}_0^{i'}$ denotes any sequence of i' words. Given a value i' we can calculate both terms by parsing the string $\mathring{w}_0^i X$, where each word $\mathring{w}$ in $\mathring{w}_0^i X$ is a special word that can freely act as any preterminal. The denominator can then be calculated by summing the forward probabilities of the last word $\mathring{w}_{i-1}^i$ as in Eq. 5, and the numerator by summing the forward probability of X, as in Eq. 4.

5.2.3 Prior: the denominator

The denominator in the calculation of the prior can be calculated in a way analogous to the numerator's second term (Section 5.2.2):

$$\begin{aligned} P(w_{i'}^j|E_0) &= \frac{P(w_{i'}^j,E_0)}{P(E_0)} \\ &= \frac{P(S \stackrel{*}{\Rightarrow} \mathring{w}_0^{i'} w_{i'}^j \mu)}{P(S \stackrel{*}{\Rightarrow} \mathring{w}_0^{i'} \mu)} \end{aligned}$$

5.2.4 Prior: starting position probability

Finally, we must calculate the second term in Eq. 8, the probability of the starting position $P(i=i')$. Given that all our terms are conditional on the existence of all words in the sentence up to i' (E_0), the probability of a starting position $P(i)$ is the probability of drawing i' randomly from the set of positions in sentences generated by the grammar such that all words up to that position exist. For most language grammars, this distribution can be easily approximated by a sample of sentences generated from the SCFG, since most of the probability mass is concentrated in small indices.

6 Experiment

We tested the predictions of an implemented version of our model on the materials from Tabor et al. (2004). To generate quantitative predictions, we created a small grammar of relevant syntactic rules, and estimated the rule probabilities from syntactically annotated text. We calculated summed K-L divergence of the prior from the posterior vector for each word in the Tabor et al. items, and predict this sum to be largest at the critical region when the sentence has an effect of local coherence.

6.1 Methods

6.1.1 Grammar

We defined a small SCFG for the problem, and estimated its rule probabilities using the parsed Brown corpus. The resulting SCFG is identical to that used in Levy (2008b) and is given in Table 2.

6.1.2 Lexicon

Lexical rewrite probabilities for part-of-speech tags were also estimated using the entire parsed Brown corpus.

6.1.3 Materials

The materials were taken from Experiment 1 of Tabor et al. (2004). We removed 8 of their 20 items for which our trained model either did not know the critical verb or did not know the syntactic structure of some part of the sentence. For the other 12 items, we replaced unknown nouns (9 instances) and unknown non-critical verbs (2 instances), changed one plural noun to singular, and dropped one sentence-initial prepositional phrase.

Table 2: The SCFG used in Experiment 3. Rule weights given as negative log-probabilities in bits.

Rule		Weight
ROOT	→ S	0
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 NN NN	4.7
NP-base	→ DT NN	1.9
NP-base	→ DT JJ NN	3.8
NP-base	→ PRP	1.0
NP-base	→ NNP	3.1
VP/NP	→ VBD NP	4.0
VP/NP	→ VBD	0.1
VP	→ VBD PP	2.0
VP	→ VBD NP	0.7
VP	→ VBD	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

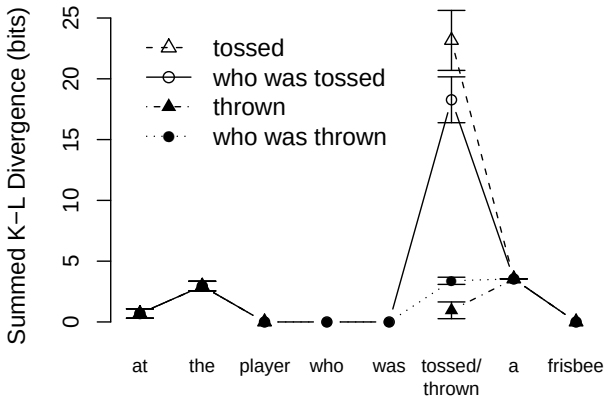

Figure 2: Summed K-L divergence of the prior from the posterior vectors at each word: Substring length 1

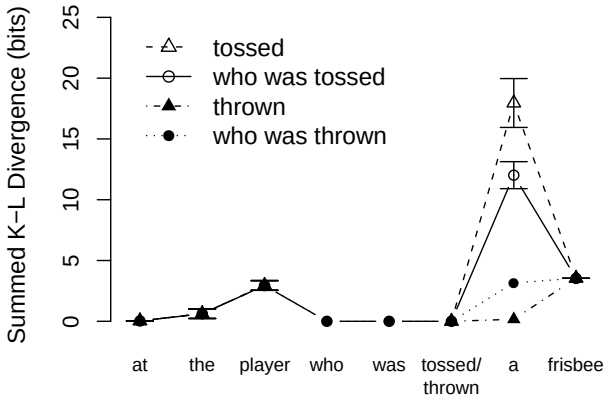

Figure 3: Summed K-L divergence of the prior from the posterior vectors at each word: Substring length 2

6.2 Procedure

For these 12 items, we ran our model on the four conditions in (1). For each word, we calculated the prior and posterior vectors for substrings of three lengths at w_i. The summed K-L divergence is reported for a substring length of 1 word using a prior of $P(X_{i-1}^{k\geq i}|w_{i-1}^{i})$, for a length of 2 using $P(X_{i-2}^{k\geq i}|w_{i-2}^{i})$, and for a length of 3 using $P(X_{i-3}^{k\geq i}|w_{i-3}^{i})$. For all lengths, we predict the summed divergence to be greater at critical words for the part-of-speech ambiguous conditions (1a,1b) than for unambiguous (1c,1d), because the part-of-speech unambiguous verbs cannot give rise to a prior that predicts for a sentence to begin. For a substring length of 3, we also predict that the divergence is superadditively greatest in the ambiguous reduced condition (1a), because of the possibility of starting a sentence with *the player tossed*.

6.3 Results

The results of the experiment are shown in Figures 2–4. For all three substring lengths, the model predicts difficulty to be greater in the ambiguous conditions at the critical words (*tossed/thrown a frisbee*). For 1-word substrings, the effect is localized on the critical verb (*tossed/thrown*), for 2-word substrings it is localized on the word directly following the critical verb (*tossed/thrown a*), and for 3-word substrings there are two effects: one on the critical verb (*the player tossed/thrown*) and one two words later (*tossed/thrown a frisbee*). Furthermore, for 3-word substrings, the effect is superadditively greatest for *the player tossed*. These results thus nicely confirm

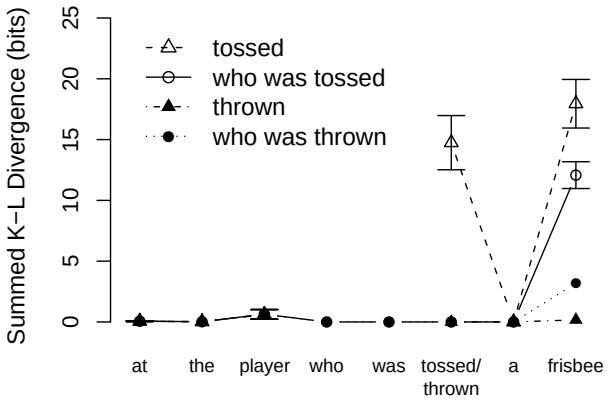

Figure 4: Summed K-L divergence of the prior from the posterior vectors at each word: Substring length 3

both of our predictions and demonstrate that a model in which large belief updates from a bottom-up prior to a posterior induce difficulty is capable of accounting for effects of local coherences.

7 Conclusion

This paper has described a model of local coherence effects in sentence processing, which views the process of integrating a string of words w_i^j into a sentence as a process of updating prior beliefs about the structures spanning those words to posterior beliefs. These prior beliefs are simply the probabilities of those structures given only the words being integrated, and the posterior beliefs are the probabilities given the entire sentence processed thus far. Difficulty is predicted to result whenever this update is large – which we model in terms of a large summed K-L divergence of the prior from the posterior vector. We demonstrated a method of normatively calculating these probabilities from probabilistic Earley parses and used this implemented model to make predictions for the materials for the original experimental result of effects of local coherences (Tabor et al., 2004). Our results demonstrated that the model predicts difficulty to occur at the correct part of the sentence in the correct condition.

We improve on existing models in two ways. First, we make predictions for where local coherences should obtain for an arbitrary SCFG, not just one particular class of sentences. This allows the model to scale up for use with a broad coverage grammar and to make predictions for arbitrary sentences, which was not possible with a model such as Tabor & Hutchins (2004).

Second, our model gives a rational basis to an effect which has typically been seen to result from irrationality of the human sentence processor. Specifically, the cost that our model describes of updating bottom-up prior beliefs to in-context posterior beliefs can be viewed as resulting from a rational process in the case that the bottom-up prior is available to the human sentence processor more rapidly than the in-context posterior. Interestingly, the fact that the prior is actually more difficult to compute than the posterior suggests that the only way it would be available more rapidly is if it is precomputed. Thus, our model provides the insight that, to the extent that comprehenders are behaving rationally in producing effects of local coherences, this may indicate that they have precomputed the likely syntactic structures of short sequences of words. While it may be unlikely that they calculate these probabilities for sequences directly from their grammar as we do in this paper, there could be a number of ways to approximate this prior: for example, given a large enough corpus, these probabilities could be approximated for any string of words that appears sufficiently often by merely tracking the structures the string has each time it occurs. Such a hypothesis for how comprehenders approximate the prior could be tested by manipulating the frequency of the relevant substrings in sentences with local coherences.

This work can be extended in a number of ways. As already mentioned, one logical step is using a broad-coverage grammar. Another possibility relates to the problem of correlations between the different components of the prior and posterior vectors. For example, in our small grammar, whenever a ROOT category begins, so does an S, an S-base, and an NP-base. Dimensionality reduction techniques on our vectors may be able to remove such correlations. These steps and more exhaustive evaluation of a variety of datasets remain for the future.

Acknowledgments

This research was supported by NIH Training Grant T32-DC000041 from the Center for Research in Language at UCSD to the first author.

References

Gerry T.M. Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73:247–264.

Jay Earley. 1970. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, July. Association for Computational Linguistics.

Dmitriy Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In Michael Collins and Mark Steedman, editors, *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 65–72, Sapporo, Japan. Association for Computational Linguistics.

Edward Gibson. 2006. The interaction of top-down and bottom-up statistics in the resolution of syntactic category ambiguity. *Journal of Memory and Language*, 54:363–388.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 159–166, New Brunswick, NJ. Associate for Computational Linguistics.

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

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 317–324, Barcelona, Spain, July. Association for Computational Linguistics.

Lars Konieczny and Daniel Müller. 2007. Local coherence interpretation in written and spoken language. Presented at the 20th Annual CUNY Conference on Human Sentence Processing. La Jolla, CA.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 849–856, Cambridge, MA. MIT Press.

Roger Levy, Florencia Reali, and Thomas L. Griffiths. 2009. Modeling the effects of memory on human online sentence processing with particle filters. In *Proceedings of NIPS*.

Roger Levy. 2008a. Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177.

Roger Levy. 2008b. A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Honolulu, Hawaii, October. Association for Computational Linguistics.

Srini Narayanan and Daniel Jurafsky. 2001. A Bayesian model predicts human parse preference and reading time in sentence processing. In T.G. Dietterich, S Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 59–65, Cambridge, MA. MIT Press.

Andreas Stolcke. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201.

Whitney Tabor and Sean Hutchins. 2004. Evidence for self-organized sentence processing: Digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):431–450.

Whitney Tabor, Bruno Galantucci, and Daniel Richardson. 2004. Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50:355–370.

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