Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing

Brian Roark[†] **Asaf Bachrach**[‡] **Carlos Cardenas**[◦] **and Christophe Pallier**[‡]
[†]Center for Spoken Language Understanding, Oregon Health & Science University
[‡] INSERM-CEA Cognitive Neuroimaging Unit, Gif sur Yvette, France [◦]MIT
roark@cslu.ogi.edu asafbac@gmail.com cardenas@mit.edu christophe@pallier.org

Abstract

A number of recent publications have made use of the incremental output of stochastic parsers to derive measures of high utility for psycholinguistic modeling, following the work of Hale (2001; 2003; 2006). In this paper, we present novel methods for calculating separate lexical and syntactic surprisal measures from a single incremental parser using a lexicalized PCFG. We also present an approximation to entropy measures that would otherwise be intractable to calculate for a grammar of that size. Empirical results demonstrate the utility of our methods in predicting human reading times.

1 Introduction

Assessment of linguistic complexity has played an important role in psycholinguistics and neurolinguistics for a long time, from the use of mean length of utterance and related scores in child language development (Klee and Fitzgerald, 1985), to complexity scores related to reading difficulty in human sentence processing studies (Yngve, 1960; Frazier, 1985; Gibson, 1998). Operationally, such linguistic complexity scores are derived via deterministic manual (human) annotation and scoring algorithms of language samples. Natural language processing has been employed to automate the extraction of such measures (Sagae et al., 2005; Roark et al., 2007), which can have high utility in terms of reduction of time required to annotate and score samples. More interestingly, however, novel data driven methods are being increasingly employed in this sphere, yielding language sample characterizations that *require* NLP in their derivation. For example, scores derived from variously estimated language models have been used to evaluate and classify language samples associated with neurodevelopmental or neurodegenerative disorders (Roark et al., 2007; Solorio and Liu, 2008; Gabani et al., 2009), as well as within general studies of human sentence processing (Hale, 2001; 2003; 2006). These scores cannot feasibly be derived by hand, but rather rely on large-scale statistical models and structured inference algorithms to be derived. This is quickly becoming an important application of NLP, making possible new methods in the study of human language processing in both typical and impaired populations.

The use of broad-coverage parsing for psycholinguistic modeling has become very popular recently. Hale (2001) suggested a measure (surprisal) derived from an Earley (1970) parser using a probabilistic context-free grammar (PCFG) for psycholinguistic modeling; and in later work (Hale, 2003; 2006) he suggested an alternate parser-derived measure (entropy reduction) that may also account for some human sentence processing performance. Recent work continues to advocate surprisal in particular as a very useful measure for predicting processing difficulty (Boston et al., 2008a; Boston et al., 2008b; Demberg and Keller, 2008; Levy, 2008), and the measure has been derived using a variety of incremental (left-to-right) parsing strategies, including an Earley parser (Boston et al., 2008a), the Roark (2001) incremental top-down parser (Demberg and Keller, 2008), and an n-best version of the Nivre et al. (2007) incremental dependency parser (Boston et al., 2008a; 2008b). Deriving such measures by hand, even for a relatively limited set of stimuli, is not feasible, hence parsing plays a critical role in this developing psycholinguistic enterprise.

There is no single measure that can account for all of the factors influencing human sentence processing performance, and some of the most recent work on using parser-derived measures for psycholinguistic modeling has looked to try to derive multiple, complementary measures. One of

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 324–333,
Singapore, 6-7 August 2009.

the key distinctions being looked at is syntactic versus lexical expectations (Gibson, 2006). For example, in Demberg and Keller (2008), trials were run deriving surprisal from the Roark (2001) parser under two different conditions: fully lexicalized parsing, and fully unlexicalized parsing (to pre-terminal part-of-speech tags). Boston et al. (2008a) capture a similar distinction by making use of an unlexicalized PCFG within an Earley parser and a fully lexicalized unlabeled dependency parser (Nivre et al., 2007). As Demberg and Keller (2008) point out, fully unlexicalized grammars ignore important lexico-syntactic information when deriving the "syntactic" expectations, such as subcategorization preferences of particular verbs, which are generally accepted to impact syntactic expectations in human sentence processing (Garnsey et al., 1997). Demberg and Keller argue, based on their results, for unlexicalized surprisal instead of lexicalized surprisal. Here we present a novel method for deriving separate syntactic and lexical surprisal measures from a fully lexicalized incremental parser, to allow for rich probabilistic grammars to be used to derive either measure, and demonstrate the utility of this method versus that of Demberg and Keller in empirical trials.

The use of large-scale lexicalized grammars presents a problem for using an Earley parser to derive surprisal or for the calculation of entropy as Hale (2003; 2006) defines it, because both methods require matrix inversion of a matrix with dimensionality the size of the non-terminal set. With very large lexicalized PCFGs, the size of the non-terminal set is too large for tractable matrix inversion. The use of an incremental, beam-search parser provides a tractable approximation to both measures. Incremental top-down and left-corner parsers have been shown to effectively (and efficiently) make use of non-local features from the left-context to yield very high accuracy syntactic parses (Roark, 2001; Henderson, 2003; Collins and Roark, 2004), and we will use such rich models to derive our scores.

In addition to teasing apart syntactic and lexical surprisal (defined explicitly in §3), we present an approximation to the full entropy that Hale (2003; 2006) used to define the entropy reduction hypothesis. Such an entropy measure is derived via a predictive step, advancing the parses independently of the input, as described in §3.3. We also present syntactic and lexical alternatives for this measure, and demonstrate the utility of making such a distinction for entropy as well as surprisal.

The purpose of this paper is threefold. First, to present a careful and well-motivated decomposition of lexical and syntactic expectation-based measures from a given lexicalized PCFG. Second, to explicitly document methods for calculating these and other measures from a specific incremental parser. And finally, to present some empirical validation of the novel measures from real reading time trials. We modified the Roark (2001) parser to calculate the discussed measures[1], and the empirical results in §4 show several things, including: 1) using a fully lexicalized parser to calculate syntactic surprisal and entropy provides higher predictive utility for reading times than these measures calculated via unlexicalized parsing (as in Demberg and Keller); and 2) syntactic entropy is a useful predictor of reading time.

2 Notation and preliminaries

A probabilistic context-free grammar (PCFG) $G = (V, T, S^\dagger, P, \rho)$ consists of a set of non-terminal variables V; a set of terminal items (words) T; a special start non-terminal $S^\dagger \in V$; a set of rule productions P of the form $A \rightarrow \alpha$ for $A \in V$, $\alpha \in (V \cup T)^*$; and a function ρ that assigns probabilities to each rule in P such that for any given non-terminal symbol $X \in V$, $\sum_\alpha \rho(X \rightarrow \alpha) = 1$.

For a given rule $A \rightarrow \alpha \in P$, let the function RHS return the right-hand side of the rule, i.e., $\text{RHS}(A \rightarrow \alpha) = \alpha$. Without loss of generality, we will assume that for every rule $A \rightarrow \alpha \in P$, one of two cases holds: either $\text{RHS}(A \rightarrow \alpha) \in T$ or $\text{RHS}(A \rightarrow \alpha) \in V^*$. That is, the right-hand side sequences consist of either (1) exactly one terminal item, or (2) zero or more non-terminals.

Let $W \in T^n$ be a terminal string of length n, i.e., $W = W_1 \ldots W_n$ and $|W| = n$. Let $W[i, j]$ denote the substring beginning at word W_i and ending at word W_j of the string. Then $W_{|W|}$ is the last word in the string, and $W[1, |W|]$ is the string as a whole. Adjacent strings represent concatenation, i.e., $W[1, i]W[i{+}1, j] = W[1, j]$. Thus $W[1, i]w$ represents the string where $W_{i+1} = w$.

We can define a "derives" relation (denoted $\Rightarrow_G$ for a given PCFG G) as follows: $\beta A \gamma \Rightarrow_G \beta\alpha\gamma$ if and only if $A \rightarrow \alpha \in P$. A string $W \in T^*$ is in the language of a grammar G if and only if $S^\dagger \stackrel{+}{\Rightarrow}_G W$, i.e., a sequence of one or more derivation steps yields the string from the start

[1] The parser version will be made publicly available.

non-terminal. A *leftmost* derivation begins with $S^\dagger$ and each derivation step replaces the leftmost non-terminal A in the yield with some α such that $A \rightarrow \alpha \in P$. For a leftmost derivation $S^\dagger \stackrel{*}{\Rightarrow}_G \alpha$, where $\alpha \in (V \cup T)^*$, the sequence of derivation steps that yield α can be represented as a tree, with the start symbol $S^\dagger$ at the root, and the "yield" sequence α at the leaves of the tree. A *complete* tree has only terminal items in the yield, i.e., $\alpha \in T^*$; a *partial* tree has some non-terminal items in the yield. With a leftmost derivation, the yield $\alpha = \beta\gamma$ partitions into an initial sequence of terminals $\beta \in T^*$ followed by a sequence of non-terminals $\gamma \in V^*$. For a complete derivation, $\gamma = \epsilon$; for a partial derivation $\gamma \in V^+$, i.e., one or more non-terminals. Let $\mathcal{T}(G, W[1,i])$ be the set of complete trees with $W[1,i]$ as the yield of the tree, given PCFG G.

A leftmost derivation D consists of a sequence of $|D|$ steps. Let D_i represent the i^{th} step in the derivation D, and $D[i,j]$ represent the subsequence of steps in D beginning with D_i and ending with D_j. Note that $D_{|D|}$ is the last step in the derivation, and $D[1,|D|]$ is the derivation as a whole. Each step D_i in the derivation is a rule in G, i.e., $D_i \in P$ for all i. The probability of the derivation and the corresponding tree is:

$$\rho(D) \quad = \quad \prod_{i=1}^{m} \rho(D_i) \tag{1}$$

Let $\mathcal{D}(G, W[1,i])$ be the set of all possible leftmost derivations D (with respect to G) such that $\text{RHS}(D_{|D|}) = W_i$. These are the set of partial leftmost derivations whose last step used a production with terminal W_i on the right-hand side. The prefix probability of $W[1,i]$ with respect to G is

$$\text{PrefixProb}_G(W[1,i]) \quad = \sum_{D \in \mathcal{D}(G,W[1,i])} \rho(D) \tag{2}$$

From this prefix probability, we can calculate the conditional probability of each word $w \in T$ in the terminal vocabulary, given the preceding sequence $W[1,i]$ as follows:

$$\begin{aligned} P_G(w \mid W[1,i]) &= \frac{\text{PrefixProb}_G(W[1,i]w)}{\sum_{w' \in T} \text{PrefixProb}_G(W[1,i]w')} \\ &= \frac{\text{PrefixProb}_G(W[1,i]w)}{\text{PrefixProb}_G(W[1,i])} \end{aligned} \tag{3}$$

This, in fact, is precisely the conditional probability that is used for language modeling for such applications as speech recognition and machine translation, which was the motivation for various syntactic language modeling approaches (Jelinek and Lafferty, 1991; Stolcke, 1995; Chelba and Jelinek, 1998; Roark, 2001).

As with language modeling, it is important to model the end of the string as well, usually with an explicit end symbol, e.g., </s>. For a string $W[1,i]$, we can calculate its prefix probability as shown above. To calculate its complete probability, we must sum the probabilities over the set of complete trees $\mathcal{T}(G, W[1,i])$. In such a way, we can calculate the conditional probability of ending the string with </s> given $W[1,i]$ as follows:

$$P_G(\text{</s>} \mid W[1,i]) = \frac{\sum_{D \in \mathcal{T}(G,W[1,i])} \rho(D)}{\text{PrefixProb}_G(W[1,i])} \tag{4}$$

2.1 Incremental top-down parsing

In this section, we review relevant details of the Roark (2001) incremental top-down parser, as configured for use here. As presented in Roark (2004), the probabilities in the PCFG are smoothed so that the parser is guaranteed not to fail due to garden pathing, despite following a beam search strategy. Hence there is always a non-zero prefix probability as defined in Eq. 2.

The parser follows a top-down leftmost derivation strategy. The grammar is factored so that every production has either a single terminal item on the right-hand side or is of the form A $\rightarrow$ B A-B, where A,B $\in V$ and the factored A-B category can expand to any sequence of children categories of A that can follow B. This factorization of n-ary productions continues to nullary factored productions, i.e., the end of the original production A $\rightarrow$ $\text{B}_1 \ldots \text{B}_n$ is signaled with an empty production A-B_1-$\ldots$-$\text{B}_n \rightarrow \epsilon$.

The parser maintains a set of possible connected derivations, weighted via the PCFG. It uses a beam search, whereby the highest scoring derivations are worked on first, and derivations that fall outside of the beam are discarded. The reader is referred to Roark (2001; 2004) for specifics about the beam search.

The model conditions the probability of each production on features extracted from the partial tree, including non-local node labels such as parents, grandparents and siblings from the left-context, as well as c-commanding lexical items. Hence this is a lexicalized grammar, though the incremental nature precludes a general head-first strategy, rather one that looks to the left-context for c-commanding lexical items.

To avoid some of the early prediction of structure, the version of the Roark parser that we used

performs an additional grammar transformation beyond the simple factorization already described – a selective left-corner transform of left-recursive productions (Johnson and Roark, 2000). In the transformed structure, slash categories are used to avoid predicting left-recursive structure until some explicit indication of modification is present, e.g., a preposition.

The final step in parsing, following the last word in the string, is to "complete" all non-terminals in the yield of the tree. All of these open non-terminals are composite factored categories, such as S-NP-VP, which are "completed" by rewriting to ϵ. The probability of these ϵ productions is what allows for the calculation of the conditional probability of ending the string, shown in Eq. 4.

One final note about the size of the non-terminal set and the intractability of exact inference for such a scenario. The non-terminal set not only includes the original atomic non-terminals of the grammar, but also any categories created by grammar factorization (S-NP) or the left-corner transform (NP/NP). Additionally, however, to remain context-free, the non-terminal set must include categories that incorporate non-local features used by the statistical model into their label, including parents, grandparents and sibling categories in the left-context, as well as c-commanding lexical heads. These non-local features must be made local by encoding them in the non-terminal labels, leading to a very large non-terminal set and intractable exact inference. Heavy smoothing is required when estimating the resulting PCFG. The benefit of such a non-terminal set is a rich model, which enables a more peaked statistical distribution around high quality syntactic structures and thus more effective pruning of the search space. The fully connected left-context produced by top-down derivation strategies provides very rich features for the stochastic parsing models. See Roark (2001; 2004) for discussion of these issues.

We now turn to measures that can be derived from the parser which may be of use for psycholinguistic modeling.

3 Parser and grammar derived measures

3.1 Surprisal

The *surprisal* at word W_i is the negative log probability of W_i given the preceding words. Using prefix probabilities, this can be calculated as:

$$S_G(W_i) \;=\; -\log \frac{\text{PrefixProb}_G(W[1,i])}{\text{PrefixProb}_G(W[1,i-1])} \quad (5)$$

Substituting equation 2 into this, we get

$$S_G(W_i) \;=\; -\log \frac{\sum_{D \in \mathcal{D}(G,W[1,i])} \rho(D)}{\sum_{D \in \mathcal{D}(G,W[1,i-1])} \rho(D)} \quad (6)$$

If we are using a beam-search parser, some of the derivations are pruned away. Let $\mathcal{B}(G, W[1,i]) \subseteq \mathcal{D}(G, W[1,i])$ be the set of derivations in the beam. Then the surprisal can be approximated as

$$S_G(W_i) \;\approx\; -\log \frac{\sum_{D \in \mathcal{B}(G,W[1,i])} \rho(D)}{\sum_{D \in \mathcal{B}(G,W[1,i-1])} \rho(D)} \quad (7)$$

Any pruning in the beam search will result in a *deficient* probability distribution, i.e., a distribution that sums to less than 1. Roark's thesis (2001) showed that the amount of probability mass lost for this particular approach is very low, hence this provides a very tight bound on the actual surprisal given the model.

3.2 Lexical and Syntactic surprisal

High surprisal scores result when the prefix probability at word W_i is low relative to the prefix probability at word W_{i-1}. Sometimes this is due to the identity of W_i, i.e., it is a surprising word given the context. Other times, it may not be the lexical identity of the word so much as the syntactic structure that must be created to integrate the word into the derivations. One would like to tease surprisal apart into "syntactic surprisal" versus "lexical surprisal", which would capture this intuition of the lexical versus syntactic dimensions to the score. Our solution to this has the beneficial property of producing two scores whose sum equals the original surprisal score.

The original surprisal score is calculated via sets of partial derivations at the point when each word W_i is integrated into the syntactic structure, $\mathcal{D}(G, W[1,i])$. We then calculate the ratio from point to point in sequence. To tease apart the lexical and syntactic surprisal, we will consider sets of partial derivations *immediately before* each word W_i is integrated into the syntactic structure, i.e., $D[1, |D|-1]$ for $D \in \mathcal{D}(G, W[1,i])$. Recall that the last derivation move for every derivation in the set is from the POS-tag to the lexical item. Hence the sequence of derivation moves that excludes the last one includes all structure except the word W_i. Then the syntactic surprisal is calculated as:

$$\text{Syn}S_G(W_i) = -\log \frac{\sum_{D \in \mathcal{D}(G,W[1,i])} \rho(D[1,|D|-1])}{\sum_{D \in \mathcal{D}(G,W[1,i-1])} \rho(D)} \quad (8)$$

and the lexical surprisal is calculated as:

$$\text{Lex}S_G(W_i) = -\log \frac{\sum_{D \in \mathcal{D}(G,W[1,i])} \rho(D)}{\sum_{D \in \mathcal{D}(G,W[1,i])} \rho(D[1,|D|-1])} \quad (9)$$

Note that the numerator of $\text{Syn}S_G(W_i)$ is the denominator of $\text{Lex}S_G(W_i)$, hence they sum to form total surprisal $S_G(W_i)$. As with total surprisal, these measures can be defined either for the full set $\mathcal{D}(G, W[1,i])$ or for a pruned beam of derivations $\mathcal{B}(G, W[1,i]) \subseteq \mathcal{D}(G, W[1,i])$.

Finally, we replicated the Demberg and Keller (2008) "unlexicalized" surprisal by replacing every lexical item in the training corpus with its POS-tag, and then parsing the POS-tags of the language samples rather than the words. This differs from our syntactic surprisal by having no lexical conditioning events for rule probabilities, and by having no ambiguity about the POS-tag of the lexical items in the string. We will refer to the resulting surprisal measure as "POS surprisal" to distinguish it from our syntactic surprisal measure.

3.3 Entropy

Entropy scores of the sort advocated by Hale (2003; 2006) involve calculation over the set of complete derivations consistent with the set of partial derivations. Hale performs this calculation efficiently via matrix inversion, which explains the use of relatively small-scale grammars with tractably sized non-terminal sets. Such methods are not tractable for the kinds of richly conditioned, large-scale PCFGs that we advocate using here. At each word in the string, the Roark (2001) top-down parser provides access to the weighted set of partial analyses in the beam; the set of complete derivations consistent with these is not immediately accessible, hence additional work is required to calculate such measures.

Let $H(\mathcal{D})$ be the entropy over a set of derivations $\mathcal{D}$, calculated as follows:

$$H(\mathcal{D}) = -\sum_{D \in \mathcal{D}} \frac{\rho(D)}{\sum_{D' \in \mathcal{D}} \rho(D')} \log \frac{\rho(D)}{\sum_{D' \in \mathcal{D}} \rho(D')} \quad (10)$$

If the set of derivations $\mathcal{D} = \mathcal{D}(G, W[1,i])$ is a set of partial derivations for string $W[1,i]$, then $H(\mathcal{D})$ is a measure of uncertainty over the partial derivations, i.e., the uncertainty regarding the correct analysis of what has already been processed. This can be calculated directly from the existing parser operations. If the set of derivations are the complete derivations *consistent* with the set of partial derivations – complete derivations that could occur over the set of possible continuations of the string – then this is a measure of the uncertainty about what is yet to come. We would like measures that can capture this distinction between (a) uncertainty of what has already been processed ("current ambiguity") versus (b) uncertainty of what is yet to be processed ("predictive entropy"). In addition, as with surprisal, we would like to tease apart the syntactic uncertainty versus lexical uncertainty.

To calculate the predictive entropy after word sequence $W[1,i]$, we modify the parser as follows: the parser extends the set of partial derivations to include all possible next words (the entire vocabulary plus </s>), and calculates the entropy over that set. This measure is calculated from just one additional word beyond the current word, and hence is an approximation to Hale's conditional entropy of grammatical continuations, which is over complete derivations. We will denote this as $H^1_G(W[1,i])$ and calculate it as follows:

$$H^1_G(W[1,i]) = H(\bigcup_{w \in T \cup \{\text{</s>}\}} \mathcal{D}(G, W[1,i]w)) \quad (11)$$

This is performing a predictive step that the baseline parser does not perform, extending the parses to all possible next words.

Unlike surprisal, entropy does not decompose straightforwardly into syntactic and lexical components that sum to the original composite measure. To tease apart entropy due to syntactic uncertainty versus that due to lexical uncertainty, we can define the set of derivations up to the preterminal (POS-tag) non-terminals as follows. Let $S(\mathcal{D}) = \{D[1,|D|-1] : D \in \mathcal{D}\}$, i.e., the set of derivations achieved by removing the last step of all derivations in $\mathcal{D}$. Then we can calculate a "syntactic" H^1_G as follows:

$$\text{Syn}H^1_G(W[1,i]) = H(\bigcup_{w \in T \cup \{\text{</s>}\}} S(\mathcal{D}(G, W[1,i]w))) \quad (12)$$

Finally, "lexical" H^1_G is defined in terms of the conditional probabilities derived from prefix probabilities as defined in Eq. 3.

$$\text{Lex}H^1_G(W[1,i]) = -\sum_{w \in T \cup \{\text{</s>}\}} P_G(w \mid W[1,i]) \log P_G(w \mid W[1,i]) \quad (13)$$

As a practical matter, these values are calculated within the Roark parser as follows. A "dummy" word is created that can be assigned every POS-tag, and the parser extends from the current state to this dummy word. (The beam threshold is greatly

expanded to allow for many possible extensions.) Then every word in the vocabulary is substituted for the word, and the appropriate probabilities calculated over the beam. Finally, the actual next word is substituted, the beam threshold is reduced to the actual working threshold, and the requisite number of analyses are advanced to continue parsing the string. This represents a significant amount of additional work for the parser – particularly for vocabulary sizes that we currently use, on the order of tens of thousands of words.

As with surprisal, we can calculate an "unlexicalized" version of the measure by training and parsing just to POS-tags. We will refer to this sort of entropy as "POS entropy".

4 Empirical validation

4.1 Subjects and stimuli

In order to test the psycholinguistic relevance of the different measures produced by the parser, we conducted a word by word reading experiment. 23 native speakers of English read 4 short texts (mean length: 883.5 words, 49.25 sentences). The texts were the written versions of narratives used in a parallel fMRI experiment making use of the same parser derived measures and whose results will be published in a different paper (Bachrach et al., 2009). The narratives contained a high density of syntactically complex structures (in the form of sentential embeddings, relative clauses and other non-local dependencies) but were constructed so as to appear highly natural. The modified version of the Roark parser, trained on the Brown Corpus section of the Penn Treebank (Marcus et al., 1993), was used to parse the different narratives and produce the word by word measures.

4.2 Procedure

Each narrative was presented line by line (certain sentences required more than one line) on a computer screen (Dell Optiplex 755 running Windows XP Professional) using Linger 2.88[2]. Each line contained 11.5 words on average. Each word would appear in its relative position on the screen. The subject would then be required to push a keyboard button to advance to the next word. The original word would then disappear and the following word appear in the subsequent position on the screen. After certain sentences a comprehension question would appear on the screen (10 per narrative). This was done in order to encourage subjects to pay attention and to provide data for a post-hoc evaluation of comprehension. After each narrative, subjects were instructed to take a short break (2 minutes on average).

4.3 Data analysis

The log (base 10) of the reaction times were analyzed using a linear mixed effects regression analysis implemented in the language R (Bates et al., 2008). Reaction times longer than 1500 ms and shorter than 150 ms (raw) were excluded from the analysis (4.8% of total data). Since button press latencies inferior to 150 ms must have been planned prior to the presentation of the word, we considered that they could not reflect stimulus driven effects. Data from the first and last words on each line were discarded.

The combined data from the 4 narratives was first modeled using a model which included order of word in the narrative[3], word length, parser-derived lexical surprisal, unigram frequency, bigram probability, syntactic surprisal, lexical entropy, syntactic entropy and mean number of parser derivation steps as numeric regressors. We also included the unlexicalized POS variants of syntactic surprisal and entropy, along the lines of Demberg and Keller (2008), as detailed in § 3. Table 1 presents the correlations between these mean-centered measures.

In addition, we modeled word class (open/closed) as a categorical factor in order to assess interaction between class and the variables of interest, since such an interaction has been observed in the case of frequency (Bradley, 1983). Finally, the random effect part of the model included intercepts for subjects, words and sentences. We report significant effects at the threshold $p < .05$.

Given the presence of significant interactions between lexical class (open/closed) and a number of the variables of interests, we decided to split the data set into open and closed class words and model these separately (linear mixed effects with the same numeric variables as in the full model).

In order to evaluate the usefulness of splitting total surprisal into lexical and syntactic components we compared, using a likelihood ratio test, a model where lexical and syntactic surprisal are modeled as distinct regressors to a model where a single regressor equal to their sum (total surprisal)

[2] http://tedlab.mit.edu/∼dr/Linger/readme.html

[3] This is a regressor to control for the trend of subjects to read faster later in the narrative.

Predictor	SynH	LexH	SynS	LexS	Freq	Bgrm	PosS	PosH	Step	WLen
Syntactic Entropy (SynH)	1.00	-0.26	0.00	0.24	-0.24	0.20	0.02	0.55	-0.05	0.18
Lexical Entropy (LexH)	-0.26	1.00	0.01	-0.40	0.43	-0.38	-0.03	0.02	0.11	-0.29
Syntactic Surprisal (SynS)	0.00	0.01	1.00	-0.12	0.08	0.18	0.77	0.21	0.38	-0.03
Lexical Surprisal (LexS)	0.24	-0.40	-0.12	1.00	-0.81	0.87	-0.10	-0.20	-0.35	0.64
Unigram Frequency (Freq)	-0.24	0.43	0.08	-0.81	1.00	-0.69	0.02	0.18	0.31	-0.72
Bigram Probability (Bgrm)	0.20	-0.38	0.18	0.87	-0.69	1.00	0.11	-0.11	-0.16	0.56
POS Surprisal (PosS)	0.02	-0.03	0.77	-0.10	0.02	0.11	1.00	0.22	0.32	0.02
POS Entropy (PosH)	0.55	0.02	0.21	-0.20	0.18	-0.11	0.22	1.00	0.16	-0.11
Derivation steps (Step)	-0.05	0.11	0.38	-0.35	0.31	-0.16	0.32	0.16	1.00	-0.24
Word Length (WLen)	0.18	-0.29	-0.03	0.64	-0.72	0.56	0.02	-0.11	-0.24	1.00

Table 1: Correlations between (mean-centered) predictors. Note that unigram frequencies were represented as logs, other scores as negative logs, hence the sign of the correlations.

was included. If the larger model provides a significantly better fit than the smaller model, this provides evidence that distinguishing between lexical and syntactic contributions to surprisal is relevant. Since total entropy is not a sum of syntactic and lexical entropy, an analogous test would not be valid in that case.

4.4 Results

All subjects successfully answered the comprehension questions (92.8% correct responses, S.D.=5.1). In the full model, we observed significant main effects of word class as well as of lexical surprisal, bigram probability, unigram frequency, syntactic entropy, POS entropy and of order in the narrative. Syntactic surprisal, lexical entropy and number of steps had no significant effect. Word length also had no significant main effect but interacted significantly with word class (open/closed). Word class also interacted significantly with lexical surprisal, unigram frequency and syntactic surprisal.

The presence of these interactions led us to construct models restricted to open and closed class items respectively. The estimated parameters are reported in Table 2. Reading time for open class words showed significant effects of unigram frequency, syntactic surprisal, syntactic entropy, POS entropy and order within the narrative. The positive effect of length approached significance. Reading time for closed class words exhibited significant effects of lexical surprisal, bigram probability, syntactic entropy and order in the narrative. Length had a non-significant negative effect, thus explaining the interaction observed in the full model.

The models with separate lexical and syntactic surprisal performed better than models including combined surprisal. For open class words, the Akaike's information criterion (AIC) was -54810 for the combined model and -54819 for the independent model (likelihood ratio test comparing the two, nested, models: $\chi^2(1)=10.7, p<.001$). For closed class items, combined model's AIC was -61467 and full model's AIC was -61469 (likelihood ratio test: $\chi^2(1)=3.54, p=0.06$).

	Estimate	Std. Error	t-value
Open-class			
(Intercept)	$2.40\times10^{+00}$	2.39×10^{-02}	100.4*
Lexical Surprisal	-1.99×10^{-04}	7.28×10^{-04}	-0.3
Word Length	8.97×10^{-04}	4.62×10^{-04}	1.9
Bigram	4.18×10^{-04}	5.27×10^{-04}	0.8
Unigram Freq	-2.43×10^{-03}	1.20×10^{-03}	-2.0*
Derivation Steps	-1.17×10^{-03}	9.02×10^{-04}	-1.3
Syntactic Entropy	2.55×10^{-03}	6.19×10^{-04}	4.1*
Lexical Entropy	3.96×10^{-04}	6.68×10^{-04}	0.6
Syntactic Surprisal	3.28×10^{-03}	9.71×10^{-04}	3.4*
Order in narrative	-1.43×10^{-05}	4.34×10^{-06}	-3.3*
POS Surprisal	-6.84×10^{-04}	8.11×10^{-04}	-0.8
POS Entropy	1.47×10^{-03}	6.05×10^{-04}	2.4*
Closed-class			
(Intercept)	$2.42\times10^{+00}$	2.32×10^{-02}	104.3*
Lexical Surprisal	2.02×10^{-03}	7.84×10^{-04}	2.6*
Word Length	-1.87×10^{-03}	1.13×10^{-03}	-1.7
Bigram	1.19×10^{-03}	4.94×10^{-04}	2.4*
Unigram Freq	1.69×10^{-03}	2.67×10^{-03}	0.6
Derivation Steps	3.01×10^{-04}	5.09×10^{-04}	0.6
Syntactic Entropy	3.15×10^{-03}	5.05×10^{-04}	6.2*
Lexical Entropy	1.83×10^{-04}	8.63×10^{-04}	0.2
Syntactic Surprisal	3.00×10^{-04}	8.35×10^{-04}	0.4
Order in narrative	-1.33×10^{-05}	3.99×10^{-06}	-3.3*
POS Surprisal	-6.46×10^{-04}	6.81×10^{-04}	-0.9
POS Entropy	6.63×10^{-04}	5.04×10^{-04}	1.3

Table 2: Estimated effects from mixed effects models on open and closed items (stars denote significance at p<.05)

4.5 Discussion

Our results demonstrate the relevance of modeling psycholinguistic processes using an incremental probabilistic parser, and the utility of the novel measures presented here. Of particular interest are: the significant effects of our syntactic entropy measure; the independent contributions of lexical surprisal, bigram probability and unigram frequency; and the differences between the predictions of the lexicalized parsing model and the unlexicalized (POS) parsing model.

The effect of entropy, or uncertainty regarding

the upcoming input independent of the surprise of that input, has been observed in non-linguistic tasks (Hyman, 1953; Bestmann et al., 2008) but to our knowledge has not been quantified before in the context of sentence processing. The usefulness of computational modeling is particularly evident in the case of entropy given the absence of any subjective procedure for its evaluation[4]. The results argue in favor of a predictive parsing architecture (Van Berkum et al., 2005). The approach to entropy here differs from the one described in Hale (2006) in a couple of ways. First, as discussed above, the calculation procedure is different – we focus on extending the derivations with just one word, rather than to all possible complete derivations. Second, and most importantly, Hale emphasizes entropy reduction (or the gain in information, given an input, regarding the rest of the sentence) as the correlate of cognitive cost while here we are interested in the amount of entropy itself (and not the size of change).

Interestingly, we observed only an effect of syntactic entropy, not lexical entropy. Recent ERP work has demonstrated that subjects do form specific lexical predictions in the context of sentence processing (Van Berkum et al., 2005; DeLong et al., 2005) and so we suspect that the absence of lexical entropy effect might be partly due to sparse data. Lexical surprisal and entropy were calculated using the internal state of a parser trained on the relatively small Brown corpus. Lexical entropy showed no significant effect while lexical surprisal affected only closed class words. This pattern of results might be due to the sparseness of the relevant information in such a small corpus (e.g., verb/object preferences) and the relevance of extra-textual dimensions (world knowledge, contextual information) to lexical-specific prediction. Closed class words are both more frequent (and hence better sampled) and are less sensitive to world knowledge, yet are often determined by the grammatical context.

Demberg and Keller (2008) made use of the same parsing architecture used here to compute a syntactic surprisal measure, but used an unlexicalized parser (down to POS-tags rather than words) for this score. Their "lexicalized" surprisal is equivalent to our total surprisal (lexical surprisal + syntactic surprisal), while their POS surprisal is derived from a completely different model. In contrast, our approach achieves lexical and syntactic measures from the same model. In order to evaluate the difference between the two approaches we added unlexicalized POS surprisal calculated along the lines of that paper to our model, along with an unlexicalized POS entropy from the same model. We found no effect of unlexicalized POS surprisal[5] and a significant (but relatively small) effect of unlexicalized POS entropy. While syntactic surprisal was correlated with POS surprisal (see Table 1) and syntactic entropy correlated with POS entropy, the fact that our syntactic measures still had a significant effect suggests that lexical information contributes towards the formation of syntactic expectations.

While the effect of surprisal calculated by an incremental top down parser has been already demonstrated (Demberg and Keller, 2008), our results argue for a distinction between the effect of lexical surprisal and that of syntactic surprisal without requiring unlexicalized parsing of the sort that Demberg and Keller advocate. It is important to keep in mind that this distinction between types of prediction (and as a consequence, prediction error) is not equivalent to the one drawn in the traditional cognitive science modularity debate, which has focused on the source of these predictions. We found a positive effect of syntactic surprisal in the case of open class words. The absence of an effect for closed class words remains to be explained.

We quantified word specific surprisal using 3 sources: the parser's internal state (lexical surprisal); probability given the preceding word (negative log bigram probability); and the unigram frequency of the word in a large corpus[6]. As can be observed in Table 1, these three measures are highly correlated[7]. This is the consequence of the smoothing in the estimation procedure but also relates to a more general fact about language use: overall, more frequent words are also words more expected to appear in a specific context (Anderson and Schooler, 1991). Despite these strong correlations, the three measures produced independent

[4]The Cloze procedure (Taylor, 1953) is one way to derive probabilities that could be used to calculate entropy, though this procedure is usually conducted with lexical elicitation, which would make syntactic entropy calculations difficult.

[5]We also ran the model including unlexicalized POS surprisal *without* our syntactic surprisal or syntactic entropy, and in this condition the unlexicalized POS surprisal measure had a nearly significant effect ($t = 1.85$), which is consistent with the results in Boston et al. (2008a) and Demberg and Keller (2008).

[6]The unigram frequencies came from the HAL corpus (Lund and Burgess, 1996). All other statistical models were estimated from the Brown Corpus.

[7]Unigram frequencies were represented as logs, the others as negative logs, hence the sign of the correlations.

effects. Unigram frequency had a significant effect for open class words while bigram probability and lexical surprisal each had an effect on reading time of closed class items. Bigram probability has been often found to affect reading time using eye movement measures. This is the first study to demonstrate an additional effect of contextual surprisal given the preceding sentential context (lexical surprisal). Demberg and Keller found no effect for surprisal once bigram and unigram probabilities were included in the model but, importantly, they did not distinguish lexical and syntactic surprisal, rather "lexicalized" and "unlexicalized" surprisal.

5 Summary

We have presented novel methods for teasing apart syntactic and lexical surprisal from a fully lexicalized parser, as well as for extending the operation of a predictive parser to capture novel entropy measures that are also shown to be relevant to psycholinguistic modeling. Such automatic methods provide psycholinguistically relevant measures that are intractable to calculate by hand. The empirical validation presented here demonstrated that the new measures – particularly syntactic entropy and syntactic surprisal – have high utility for modeling human reading time data. Our approach to calculating syntactic surprisal, based on fully lexicalized parsing, provided significant effects, while the POS-tag based (unlexicalized) surprisal – of the sort used in Boston et al. (2008a) and Demberg and Keller (2008) – did not provide a significant effect in our trials. Further, we showed an effect of lexical surprisal for closed class words even when combined with unigram and bigram probabilities in the same model. This work contributes to the important, developing enterprise of leveraging data-driven NLP approaches to derive new measures of high utility for psycholinguistic and neuropsychological studies.

Acknowledgments

Thanks to Michael Collins, John Hale and Shravan Vasishth for valuable discussions about this work. This research was supported in part by NSF Grant #BCS-0826654. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF.

References

J.R. Anderson and L.J. Schooler. 1991. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408.

A. Bachrach, B. Roark, A. Marantz, S. Whitfield-Gabrieli, C. Cardenas, and J.D.E. Gabrieli. 2009. Incremental prediction in naturalistic language processing: An fMRI study. *In preparation*.

D. Bates, M. Maechler, and B. Dai, 2008. *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-20.

S. Bestmann, L.M. Harrison, F. Blankenburg, R.B. Mars, P. Haggard, and K.J. Friston. 2008. Influence of uncertainty and surprise on human corticospinal excitability during preparation for action. *Current Biology*, 18:775–780.

M. Ferrara Boston, J.T. Hale, R. Kliegl, U. Patil, and S. Vasishth. 2008a. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1):1–12.

M. Ferrara Boston, J.T. Hale, R. Kliegl, and S. Vasishth. 2008b. Surprising parser actions and reading difficulty. In *Proceedings of ACL-08:HLT, Short Papers*, pages 5–8.

D.C. Bradley. 1983. *Computational Distinctions of Vocabulary Type*. Indiana University Linguistics Club, Bloomington.

C. Chelba and F. Jelinek. 1998. Exploiting syntactic structure for language modeling. In *Proceedings of ACL-COLING*, pages 225–231.

M.J. Collins and B. Roark. 2004. Incremental parsing with the perceptron algorithm. In *Proceedings of ACL*, pages 111–118.

K.A. DeLong, T.P. Urbach, and M. Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8):1117–1121.

V. Demberg and F. Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

J. Earley. 1970. An efficient context-free parsing algorithm. *Communications of the ACM*, 6(8):451–455.

L. Frazier. 1985. Syntactic complexity. In D.R. Dowty, L. Karttunen, and A.M. Zwicky, editors, *Natural Language Parsing*. Cambridge University Press, Cambridge, UK.

K. Gabani, M. Sherman, T. Solorio, and Y. Liu. 2009. A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In *Proceedings of NAACL-HLT*.

S.M. Garnsey, N.J. Pearlmutter, E. Myers, and M.A. Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37(1):58–93.

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

E. Gibson. 2006. The interaction of top-down and bottom-up statistics in the resolution of syntactic category ambiguity. *Journal of Memory and Language*, 54(3):363–388.

J.T. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd meeting of NAACL.*

J.T. Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

J.T. Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

J. Henderson. 2003. Inducing history representations for broad coverage statistical parsing. In *Proceedings of HLT-NAACL*, pages 24–31.

R. Hyman. 1953. Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology: General*, 45(3):188–96.

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

M. Johnson and B. Roark. 2000. Compact non-left-recursive grammars using the selective left-corner transform and factoring. In *Proceedings of COLING*, pages 355–361.

T. Klee and M.D. Fitzgerald. 1985. The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language*, 12:251–269.

R. Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

K. Lund and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28:203–208.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

B. Roark, M. Mitchell, and K. Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of BioNLP Workshop at ACL*, pages 1–8.

B. Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

B. Roark. 2004. Robust garden path parsing. *Natural Language Engineering*, 10(1):1–24.

K. Sagae, A. Lavie, and B. MacWhinney. 2005. Automatic measurement of syntactic development in child langugage. In *Proceedings of ACL*, pages 197–204.

T. Solorio and Y. Liu. 2008. Using language models to identify language impairment in Spanish-English bilingual children. In *Proceedings of BioNLP Workshop at ACL*, pages 116–117.

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

W.L. Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

J.J.A. Van Berkum, C.M. Brown, P. Zwitserlood, V.Kooijman, and P. Hagoort. 2005. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Learning and Memory*, 31(3):443–467.

V.H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.