Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing

Victoria Fossum and Roger Levy
Department of Linguistics
University of California, San Diego
9500 Gilman Dr.
La Jolla, CA 92093
{vfossum,rlevy}@ucsd.edu

Abstract

Experimental evidence demonstrates that syntactic structure influences human online sentence processing behavior. Despite this evidence, open questions remain: which type of syntactic structure best explains observed behavior–hierarchical or sequential, and lexicalized or unlexicalized? Recently, Frank and Bod (2011) find that unlexicalized sequential models predict reading times better than unlexicalized hierarchical models, relative to a baseline prediction model that takes word-level factors into account. They conclude that the human parser is insensitive to hierarchical syntactic structure. We investigate these claims and find a picture more complicated than the one they present. First, we show that incorporating additional lexical n-gram probabilities estimated from several different corpora into the baseline model of Frank and Bod (2011) eliminates all differences in accuracy between those unlexicalized sequential and hierarchical models. Second, we show that lexicalizing the hierarchical models used in Frank and Bod (2011) significantly improves prediction accuracy relative to the unlexicalized versions. Third, we show that using state-of-the-art lexicalized hierarchical models further improves prediction accuracy. Our results demonstrate that the claim of Frank and Bod (2011) that sequential models predict reading times better than hierarchical models is premature, and also that lexicalization matters for prediction accuracy.

1 Introduction

Various factors influence human reading times during online sentence processing, including word-level factors such as word length, unigram and bigram probabilities, and position in the sentence. Yet word-level factors cannot explain many observed processing phenomena; ample experimental evidence exists for the influence of syntax on human behavior during online sentence processing, beyond what can be predicted using word-level factors alone. Examples include the English subject/object relative clause asymmetry (Gibson et al., 2005; King and Just, 1991) and anti-locality effects in German (Konieczny, 2000; Konieczny and Döring, 2003), Hindi (Vasishth and Lewis, 2006), and Japanese (Nakatani and Gibson, 2008). Levy (2008) shows that these processing phenomena can be explained by surprisal theory under a hierarchical probabilistic context-free grammar (PCFG). Other evidence of syntactic expectation in sentence processing includes the facilitation of processing at "or" following "either" (Staub and Clifton, 2006); expectations of heavy noun phrase shifts (Staub et al., 2006); ellipsis processing (Lau et al., 2006); and syntactic priming (Sturt et al., 2010).

Experimental evidence for the influence of syntax on human behavior is not limited to experiments carefully designed to isolate a particular processing phenomenon. Several broad-coverage experimental studies have shown that surprisal under hierarchical syntactic models predicts human processing difficulty on large corpora of naturally occurring text, even after word-level factors have been taken into

In: R. Levy & D. Reitter (Eds.), *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 61–69,
Montréal, Canada, June 7, 2012.

account (Boston et al., 2008; Demberg and Keller, 2008; Roark et al., 2009).

Despite this evidence, in recent work Frank and Bod (2011) challenge the notion that hierarchical syntactic structure is strictly necessary to predict reading times. They compare per-word surprisal predictions from unlexicalized hierarchical and sequential models of syntactic structure along two axes: *linguistic accuracy* (how well the model predicts the test corpus) and *psychological accuracy* (how well the model predicts observed reading times on the test corpus). They find that, while hierarchical phrase-structure grammars (PSG's) achieve better linguistic accuracy, sequential echo state networks (ESN's) achieve better psychological accuracy on the English Dundee corpus (Kennedy and Pynte, 2005). Frank and Bod (2011) do not include lexicalized syntactic models in the comparison on the grounds that, once word-level factors have been included as control predictors in the reading times model, lexicalized syntactic models do not predict reading times better than unlexicalized syntactic models (Demberg and Keller, 2008). Based on the results of their comparisons between unlexicalized models, they conclude that the human parser is insensitive to hierarchical syntactic structure.

In light of the existing evidence that hierarchical syntax influences human sentence processing, the claim of Frank and Bod (2011) is surprising. In this work, we investigate this claim, and find a picture more complicated than the one they present. We first replicate the results of Frank and Bod (2011) using the dataset provided by the authors, verifying that we obtain the same linguistic and psychological accuracies reported by the authors. We then extend their work in several ways. First, we repeat their comparisons using additional, more robustly estimated lexical n-gram probabilities as control predictors in the baseline model.[1] We show that when these additional lexical n-gram probabilities are used as control predictors, any differences in psychological accuracy between the hierarchical and sequential models used in Frank and Bod (2011) vanish. Second, while they restrict their comparisons to unlexicalized models over part-of-speech (POS) tags, we investigate the lexicalized versions of each hierarchical model, and show that lexicalization significantly improves psychological accuracy. Third, while they explore only a subset of the PSG's implemented under the incremental parser of Roark (2001), we explore a state-of-the-art lexicalized hierarchical model that conditions on richer contexts, and show that this model performs still better. Our findings demonstrate that Frank and Bod (2011)'s strong claim that sequential models predict reading times better than hierarchical models is premature, and also that lexicalization improves the psychological accuracy of hierarchical models.

2 Related Work

Several broad-coverage experimental studies demonstrate that surprisal under a hierarchical syntactic model predicts human processing difficulty on a corpus of naturally occurring text, even after word-level factors have been taken into account. Under surprisal theory (Hale, 2001; Levy, 2008), processing difficulty at word w_i is proportional to reading time at w_i, which in turn is proportional to the surprisal of w_i in the context in which it is observed: $surprisal(w_i) = -log(pr(w_i|context))$. Typically, $context \approx w_1...w_{i-1}$. Computing $surprisal(w_i)$ thus reduces to computing $-log(pr(w_i|w_1...wi-1))$. Henceforth, we refer to this original formulation of surprisal as *total surprisal*.

Boston et al. (2008) show that surprisal estimates from a lexicalized dependency parser (Nivre, 2006) and an unlexicalized PCFG are significant predictors of reading times on the German Potsdam Corpus. Demberg and Keller (2008) propose to isolate syntactic surprisal from total surprisal by replacing each word with its POS tag, then calculating surprisal as usual under the incremental probabilistic phrase-structure parser of Roark (2001). (Following Roark et al. (2009), we hereafter refer to this type of surprisal as *POS surprisal*.) They find that only POS surprisal, not total surprisal, is a significant predictor of reading time predictions on the English Dundee corpus.

Demberg and Keller (2008)'s definition of POS surprisal introduces two constraints. First, by omit-

[1] By *robustly estimated*, we mean that these probabilities are estimated from larger corpora and use a better smoothing method (Kneser-Ney) than the lexical n-grams of Frank and Bod (2011).

ting lexical information from the conditioning context, they ignore differences among words within a syntactic category that can influence syntactic expectations about upcoming material. Second, by replacing words with their most likely POS tags, they treat POS tags as veridical, observed input rather than marginalizing over all possible latent POS tag sequences consistent with the observed words.

Roark et al. (2009) propose a more principled way of decomposing total surprisal into its syntactic and lexical components, defining the syntactic surprisal of w_i as:

$$-log\frac{\sum_{D:yield(D)=w_1...w_i} pr(D\ minus\ last\ step)}{\sum_{D:yield(D)=w_1...w_{i-1}} pr(D)}$$

and the lexical surprisal of w_i as:

$$-log\frac{\sum_{D:yield(D)=w_1...w_i} pr(D)}{\sum_{D:yield(D)=w_1...w_i} pr(D\ minus\ last\ step)}$$

where D is the set of derivations in the parser's beam at any given point; $D : yield(D) = w_1...w_i$ is the set of all derivations in D consistent with $w_1...w_i$; and $D\ minus\ last\ step$ includes all steps in the derivation *except* for the last step, in which w_i is generated by conditioning upon all previous steps of D (including t_i).

Roark et al. (2009) show that syntactic surprisal produces more accurate reading time predictions on an English corpus than POS surprisal, and that decomposing total surprisal into its syntactic and lexical components produces more accurate reading time predictions than total surprisal taken as a single quantity. In this work, we compare not only different types of syntactic models, but also different measures of surprisal under each of those models (total, POS, syntactic-only, and lexical-only).

3 Models

Estimating $surprisal(w_i)$ amounts to calculating $-log(pr(w_i|w_1...w_{i-1}))$. Language models differ in the way they estimate the conditional probability of the event w_i given the observed context $w_1...w_{i-1}$. In the traditional formulation of surprisal under a hierarchical model, the event w_i is conditioned not only on the *observed* context $w_1...w_{i-1}$ but also on the *latent* context consisting of the syntactic trees T whose yield is $w_1...w_{i-1}$; computing $pr(w_i|w_1...w_{i-1})$ therefore requires marginalizing over all possible latent contexts T. In this formulation of surprisal, the context includes lexical information ($w_1...w_{i-1}$) as well as syntactic information ($T : yield(T) = w_1...w_{i-1}$), and the predicted event itself (w_i) contains lexical information.

Other formulations of surprisal are also possible, in which the event, observed context, and latent context are otherwise defined. In this work, we classify syntactic models as follows: *lexicalized* models include lexical information in the context, in the predicted event, or both; *unlexicalized* models include lexical information neither in the context nor in the predicted event; *hierarchical* models induce a latent context of trees compatible with the input; *sequential* models either induce no latent context at all, or induce a latent sequence of POS tags compatible with the input. Table 1 summarizes the syntactic models and various formulations of surprisal used in this work.

Following Frank and Bod (2011), we consider one type of hierarchical model (PSG's) and two types of sequential models (Markov models and ESN's).

3.1 Phrase-Structure Grammars

PSG's consists of rules expanding a parent node into children nodes in the syntactic tree, with associated probabilities. Frank and Bod (2011) use PSG's that generate POS tag sequences, not words. Under such grammars, the prefix probability of a tag sequence t is the sum of the probabilities of all trees $T : yield(T) = t_1...t_i$, where the probability of each tree T is the product of the probabilities of the rules used in the derivation of T.

Vanilla PCFG's, a special case of PSG's in which the probability of a rule depends only on the identity of the parent node, achieve sub-optimal parsing accuracy relative to grammars in which the probability of each rule depends on a richer context (Charniak, 1996; Johnson, 1998; Klein and Manning, 2003). To this end, Frank and Bod (2011) explore several variants of PSG's conditioned on successively richer contexts, including ancestor models (which condition rule expansions on ancestor nodes from 1-4 levels up in the tree) and ancestor+sibling models (which condition rule expansions on the ancestor's left sibling as well). Both sets of grammars also con-

Authors	Model	Surprisal	Observed Context	Latent Context	Predicted Event
Boston et al. (2008) Demberg and Keller (2008) Roark et al. (2009) Frank and Bod (2011) This Work	Hier.	POS	$t_i....t_{i-1}$	Trees T with yield $t_1...t_{i-1}$	t_i
Demberg and Keller (2008) Roark et al. (2009) This Work	Hier.	Total	$w_1...w_{i-1}$	Trees T with yield $t_1...t_{i-1}$	w_i
Roark et al. (2009) This Work	Hier.	Syntactic-Only	$w_1...w_{i-1}$	Trees T with yield $w_1...w_{i-1}$	t_i
Roark et al. (2009) This Work	Hier.	Lexical-Only	$w_1...w_{i-1}$	Trees T with yield $w_1...w_{i-1}$; t_i	w_i
Frank and Bod (2011) This Work	Seq.	POS	$t_i....t_{i-1}$	–	t_i
–	Seq.	Total	$w_1...w_{i-1}$	$t_1...t_{i-1}$ with yield $w_1...w_{i-1}$	w_i

Table 1: Contexts and events used to produce surprisal measures under various probabilistic syntactic models. T refers to trees; t refers to POS tags; and w refers to words.

dition rule expansions on the current head node[2].

In addition to the grammars over POS tag sequences used by Frank and Bod (2011), we evaluate PSG's over word sequences. We also include the state-of-the-art Berkeley grammar (Petrov and Klein, 2007) in our comparison. Syntactic categories in the Berkeley grammar are automatically split into fine-grained subcategories to improve the likelihood of the training corpus under the model. This increased expressivity allows the parser to achieve state-of-the-art automatic parsing accuracy, but increases grammar size considerably.[3]

3.2 Markov Models

Frank and Bod (2011) use Markov models over POS tag sequences, where the prefix probability of a sequence t is $\prod_i pr(t_i|t_{i-n+1}, t_{i-n+2}...t_{i-1})$. They use three types of smoothing: additive, Good-Turing, and Witten-Bell, and explore values of n from 1 to 3.

3.3 Echo State Networks

Unlike Markov models, ESN's (Jäger, 2001) can capture long-distance dependencies. ESN's are a type of recurrent neural network (Elman, 1991) in which only the weights from the hidden layer to the output layer are trained; the weights from the input layer to the hidden layer and from the hidden layer to itself are set randomly and do not change. In recurrent networks, the activation of the hidden layer at tag t_i depends not only on the activation of the input layer at tag t_i, but also on the activation of the hidden layer at tag t_{i-1}, which in turn depends on the activation of the hidden layer at tag t_{i-2}, and so forth. The activation of the output layer at tag t_i is therefore a function of all previous input symbols $t_1...t_{i-1}$ in the sequence. The prefix probability of a sequence t under this model is $\prod_i pr(t_i|t_1...t_{i-1})$, where $pr(t_i|t_1...t_{i-1})$ is the normalized activation of the output layer at tag t_i. Frank and Bod (2011) evaluate ESN's with 100, 200...600 hidden nodes.

4 Methods

We use two incremental parsers to calculate surprisals under the hierarchical models. For the PSG's available under the Roark et al. (2009) parser, we use that parser to calculate approximate prefix prob-

[2] or rightmost child node, if the head node is not yet available(Roark, 2001).

[3] To make parsing with the Berkeley grammar tractable under the prefix probability parser, we prune away all rules with probability less than 10^{-4}.

abilities using beam search. For the Berkeley grammar, we use a probabilistic Earley parser modified by Levy[4] to calculate exact prefix probabilities using the algorithm of Stolcke (1995). We evaluate each hierarchical model under each type of surprisal (POS, total, lexical-only, and syntactic-only), where possible.

4.1 Data Sets

Each syntactic model is trained on sections 2-21 of the Wall Street Journal (WSJ) portion of the Penn Treebank (Marcus et al., 1994), and tested on the Dundee Corpus (Kennedy and Pynte, 2005), which contains reading time measures for 10 subjects over a corpus of 2,391 sentences of naturally occurring text. Gold-standard POS tags for the Dundee corpus are obtained automatically using the Brill tagger (Brill, 1995).

Frank and Bod (2011) exclude subject/word pairs from evaluation if any of the following conditions hold true: "the word was not fixated, was presented as the first or last on a line, was attached to punctuation, contained more than one capital letter, or contained a non-letter (this included clitics)". This leaves 191,380 subject/word pairs in the data set published by Frank and Bod (2011). Because we consider lexicalized hierarchical models in addition to unlexicalized ones, we additionally exclude subject/word pairs where the word is "unknown" to the model.[5] This leaves us with a total of 148,829 subject/word pairs; all of our reported results refer to this data set.

4.2 Evaluation

Following Frank and Bod (2011), we compare the per-word surprisal predictions from hierarchical and sequential models of syntactic structure along two axes: linguistic accuracy (how well the model explains the test corpus) and psychological accuracy (how well the model explains observed reading times on the test corpus).

4.2.1 Linguistic Accuracy

Each model provides surprisal estimates $surprisal(w_i)$. The linguistic accuracy over the test corpus is $\frac{1}{n}\sum_{i=1}^{n} surprisal(w_i)$, where n is the number of words in the test corpus.

4.2.2 Psychological Accuracy

We add each model's per-word surprisal predictions to a linear mixed-effects model of first-pass reading times, then measure the improvement in reading time predictions (according to the deviance information criterion) relative to a baseline model; the resulting decrease in deviance is the psychological accuracy of the language model. Using the `lmer` package for linear mixed-effects models in R (Baayen et al., 2008), we first fit a baseline model to first-pass readings times over the test corpus. Each baseline model contains the following control predictors for each subject/word pair: `sentpos` (position of the word in the sentence), `nrchar` (number of characters in the word), `prevnonfix` (whether the previous word was fixated by the subject), `nextnonfix` (whether the next word was fixated by the subject), `logwordprob` ($log(pr(w_i))$), `logforwprob` ($log(pr(w_i|w_{i-1}))$), and `logbackprob` ($log(pr(w_i|w_{i+1}))$). When fitting each baseline model, we include all control predictors; all significant two-way interactions between them ($|t| \geq 1.96$); by-subject and by-word intercepts; and a by-subject random slope for the predictor that shows the most significant effect (`nrchar`).[6]

We evaluate the statistical significance of the difference in psychological accuracy between two predictors using a nested model comparison. If the model containing both predictors performs significantly better than the model containing only the first predictor under a χ^2 test ($p \leq 0.05$), then the second predictor accounts for variance in reading times above and beyond the first predictor, and vice versa.

[4] The prefix parser is available at: www.http://idiom.ucsd.edu/ rlevy/prefixprobabilityparser.html

[5] We consider words appearing fewer than 5 times in the training data to be unknown.

[6] In accordance with the methods of Frank and Bod (2011), "Surprisal was not included as a by-subject random slope because of the possibility that participants' sensitivity to surprisal varies more strongly for some sets of surprisal estimates than for others, making the comparisons between language models unreliable. Since subject variability is not currently of interest, it is safer to leave out random surprisal effects."

5 Results

We first replicate the results of Frank and Bod (2011) by obtaining POS surprisal values directly from the authors' published dataset for each syntactic model, then evaluating the psychological accuracy of each of those models relative to the baseline model defined above.[7]

Baseline Model with Additional Lexical N-grams Next, we explore the impact of the lexical n-gram probabilities used as control predictors upon psychological accuracy. Frank and Bod (2011) state that they compute lexical unigram and bigram probabilities via linear interpolation between estimates from the British National Corpus and the Dundee corpus itself (p.c.); upon inspection, we find that the bigram probabilities released in their published data set (which are consistent with their published experimental results) more closely resemble probabilities estimated from the Dundee corpus alone. Because of the small size of the Dundee corpus, lexical bigrams from this corpus alone are unlikely to be representative of a human's language experience.

We augment the lexical bigram probabilities used in the baseline model of Frank and Bod (2011) with additional lexical unigram and bigrams estimated using the SRILM toolkit (Stolcke, 2002) with Kneser-Ney smoothing from three corpora: sections 2-21 of the WSJ portion of the Penn Treebank, the Brown corpus, and the British National corpus. We include these additional predictors and all two-way interactions between them in the baseline model. Figure 1 shows that the relative differences in psychological accuracy between unlexicalized hierarchical and sequential models vanish under this stronger baseline condition.[8]

Unlexicalized Hierarchical Models We then calculate POS surprisal values under each of the ancestor (a1-a4) and the ancestor+sibling (s1-s4) hierarchical models ourselves, using the parser of Roark et al. (2009). We also calculate POS surprisal under the Berkeley grammar (b) using the Levy prefix probability parser. Figure 2 shows the accuracies of these models.[9]

Lexicalized Hierarchical Models Next, we lexicalize the hierarchical models. Figure 3 shows the results of computing total surprisal under each lexicalized hierarchical model (a1-a4T, s1-s4T, and bT). The lexicalized models improve significantly upon their unlexicalized counterparts ($\chi^2 = 7.52$ to $12.47, p \leq 0.01$) in all cases; by contrast, the unlexicalized models improve significantly upon their lexicalized counterparts ($\chi^2 = 4.05$ to $5.92, p \leq 0.05$) only in some cases (s1-s4). Each lexicalized model improves significantly upon e4, the best unlexicalized model of Frank and Bod (2011) ($\chi^2 = 6.96$ to $23.45, p \leq 0.01$), though e4 also achieves a smaller but still significant improvement upon each of the lexicalized models ($\chi^2 = 4.49$ to 7.58, $p \leq 0.05$). The lexicalized Berkeley grammar (bT) achieves the highest linguistic and psychological accuracy; the improvement of bT upon e4 is substantial and significant ($\chi^2(1) = 23.45$, $p \leq 0.001$), while the improvement of e4 upon bT is small but still significant ($\chi^2(1) = 4.50$, $p \leq 0.1$). Estimated coefficients for surprisal estimates under each lexicalized hierarchical model are shown in Table 2.[10]

Decomposing Total Surprisal Figure 3 shows the results of decomposing total surprisal (a1-a4T, s1-s4T) into its lexical and syntactic components, then entering both components as predictors into the mixed-effects model (a1-a4LS, s1-s4LS).[11] For each grammar, the psychological accuracy of the surprisal estimates is slightly higher when both lexical and syntactic surprisal are entered as predictors, though the differences are not statistically significant.

[7] The only difference between our results and the original results in Figure 2 of Frank and Bod (2011) is that we evaluate accuracy over a subset of the subject/items pairs used in Frank and Bod (2011) (see Section 4.1 for details).

[8] The psychological accuracies of the best sequential model (e4) and the best hierarchical model (s3) used in Frank and Bod (2011) relative to the stronger baseline with additional lexical n-grams are not significantly different, according to a χ^2 test.

[9] Our POS surprisal estimates have slightly worse linguistic accuracy but slightly better psychological accuracy than Frank and Bod (2011); these differences are likely due to differences in beam settings and in the subset of the WSJ used as training data.

[10] Each surprisal estimate predicts reading times in the expected (positive) direction.

[11] Decomposing surprisal into its lexical and syntactic components is possible with the Levy prefix probability parser as well, but requires modifications to the parser; the Roark et al. (2009) parser computes these quantities explicitly by default.

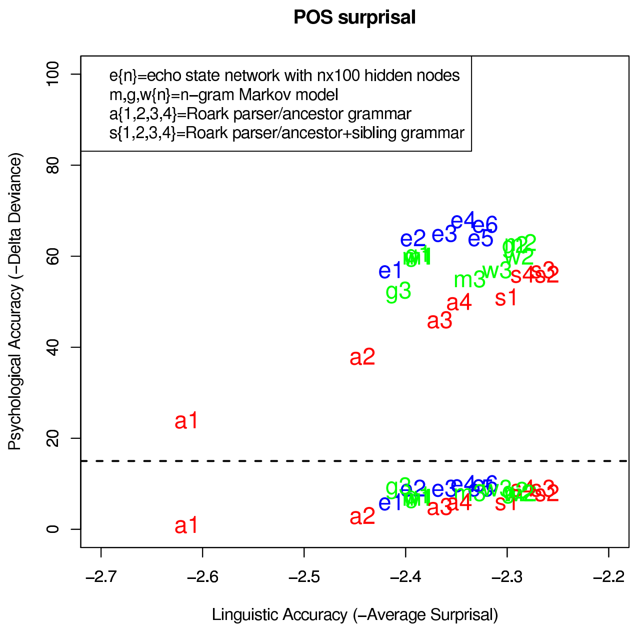

Figure 1: Psychological vs. linguistic accuracy of POS surprisal estimates from unlexicalized sequential and hierarchical models of Frank and Bod (2011) relative to baseline system of Frank and Bod (2011) (shown above dotted line), and relative to a baseline system including additional lexical unigrams and bigrams (shown below dotted line). Incorporating additional lexical n-grams into baseline system virtually eliminates all differences in psychological accuracy among models.

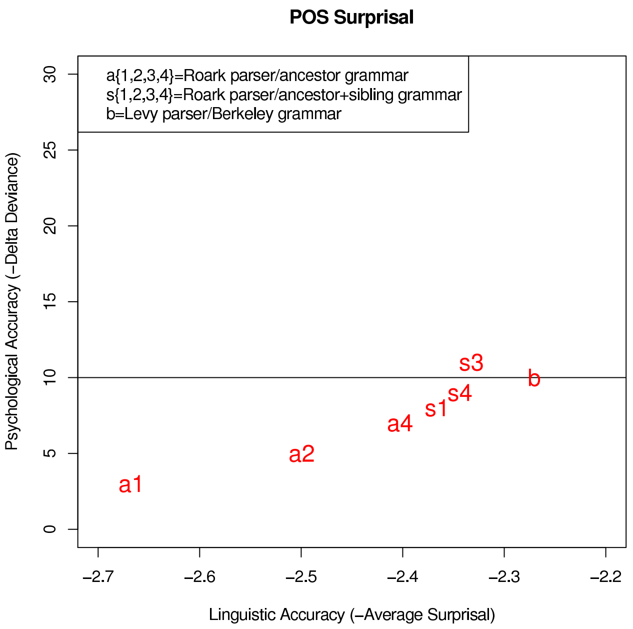

Figure 2: Psychological vs. linguistic accuracy of POS surprisal estimates from unlexicalized hierarchical models used in this work, relative to a baseline system with additional lexical unigrams and bigrams. Horizontal line indicates most psychologically accurate model of Frank and Bod (2011) for ease of comparison.

POS vs. Syntactic-only Surprisal Figures 2 and 4 show the results of computing POS surprisal (a1-a4, s1-s4) and syntactic-only surprisal (a1-a4S, s1-s4S), respectively, under each of the Roark grammars. While syntactic surprisal achieves slightly higher psychological accuracy than POS surprisal for each model, the difference is statistically significant in only one case (s1).

6 Discussion

In the presence of additional lexical n-gram control predictors, all gaps in performance between the unlexicalized sequential and hierarchical models used in Frank and Bod (2011) vanish (Figure 1). Frank and Bod (2011) do not include lexicalized hierarchical models in their study; our results indicate that lexicalizing hierarchical models improves their psychological accuracy significantly compared to the unlexicalized versions. Overall, the lexicalized hierarchical model with the highest linguistic accuracy (Berkeley) also achieves the highest psychological accuracy.

Decomposing total surprisal into its lexical- and syntactic-only components improves psychological accuracy, but this improvement is not statistically significant. Computing syntactic-only surprisal instead of POS surprisal improves psychological accuracy, but this improvement is statistically significant in only one case (s1).

7 Conclusion and Future Work

Frank and Bod (2011) claim that sequential unlexicalized syntactic models predict reading times better than hierarchical unlexicalized syntactic models, and conclude that the human parser is insensitive to hierarchical syntactic structure. We find that the picture is more complicated than this. We show, first, that the gap in psychological accuracy between the unlexicalized hierarchical and sequential models of Frank and Bod (2011) vanishes when additional,

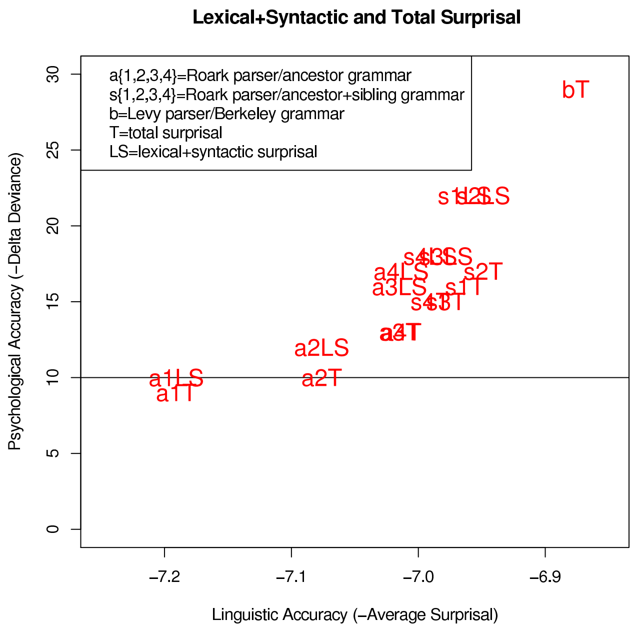

Figure 3: Psychological vs. linguistic accuracy of lexical+syntactic (LS) and total (T) surprisal estimates from lexicalized hierarchical models used in this work, relative to baseline system with additional lexical unigrams and bigrams as control predictors. Decomposing total surprisal into lexical-only and syntactic-components improves psychological accuracy. Horizontal line indicates most psychologically accurate model of (Frank and Bod, 2011).

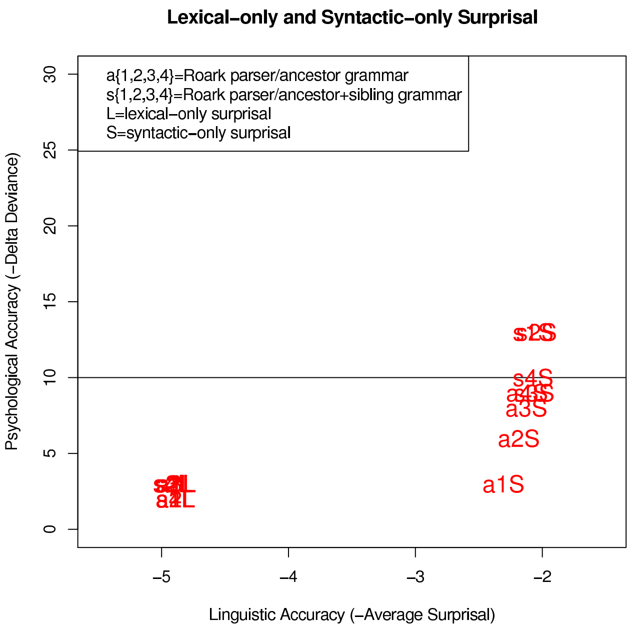

Figure 4: Psychological vs. linguistic accuracy of lexical-only (L) and syntactic-only (S) surprisal estimates from lexicalized hierarchical models used in this work, relative to baseline system with additional lexical unigrams and bigrams as control predictors. On its own, syntactic-only surprisal predicts reading times better than lexical-only surprisal. Horizontal line indicates most psychologically accurate model of (Frank and Bod, 2011).

Surprisal	Coef.	$\lvert t\rvert$	Surprisal	Coef.	$\lvert t\rvert$
a1LS	0.82	2.61	a1T	1.30	2.98
a2LS	1.01	3.24	a2T	1.38	3.19
a3LS	1.14	3.65	a3T	1.56	3.60
a4LS	1.17	3.76	a4T	1.56	3.64
s1LS	1.38	4.43	s1T	1.71	4.00
s2LS	1.37	4.44	s2T	1.75	4.16
s3LS	1.20	3.90	s3T	1.64	3.91
s4LS	1.21	3.97	s4T	1.62	3.89
bT	3.15	5.34			

Table 2: Estimated coefficients and $|t|$-values for surprisal estimates shown in Figure 3. Coefficients are estimated by adding each surprisal estimate, one at a time, to the baseline model of reading times used in Figure 3.

robustly estimated lexical n-gram probabilities are incorporated as control predictors into the baseline model of reading times. Next, we show that lexicalizing hierarchical grammars improves psychological accuracy significantly. Finally, we show that using better lexicalized hierarchical models improves psychological accuracy still further. Our results demonstrate that the claim of Frank and Bod (2011) that sequential models predict reading times better than hierarchical models is premature, and that further investigation is required.

In future work, we plan to incorporate lexical information into the sequential syntactic models used in Frank and Bod (2011) so that we can compare the hierarchical lexicalized models described here against sequential lexicalized models.

Acknowledgments

The authors thank Stefan Frank for providing the dataset of Frank and Bod (2011) and a detailed specification of their experimental configuration. This research was supported by NSF grant 0953870, NIH grant 1R01HD065829, and funding from the Army Research Laboratory's Cognition & Neuroergonomics Collaborative Technology Alliance.

References

R. H. Baayen, D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. In *Journal of Memory and Language, 59, pp. 390-412*.

Marisa Ferrara Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. In *Journal of Eye Movement Research, 2(1):1, pages 1-12*.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4).

Eugene Charniak. 1996. Tree-bank grammars. In *AAAI*.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. In *Cognition, Volume 109, Issue 2, pages 193-210*.

J.L. Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2).

Stefan Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. In *Psychological Science*.

Edward Gibson, Timothy Desmet, Daniel Grodner, Duane Watson, and Kara Ko. 2005. Reading relative clauses in english. *Cognitive Linguistics*, 16(2).

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of NAACL*.

Herbert Jäger. 2001. The" echo state" approach to analysing and training recurrent neural networks. In *Technical Report GMD 148, German National Research Center for Information Technology*.

Mark Johnson. 1998. Pcfg models of linguistic tree representations. *Computational Linguistics*, 24.

A. Kennedy and J. Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision research*, 45(2).

Jonathan King and Marcel Just. 1991. Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, 30(5).

Dan Klein and Chris Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*.

Lars Konieczny and Philipp Döring. 2003. Anticipation of clause-final heads: Evidence from eye-tracking and srns. In *Proceedings of ICCS/ASCS*.

Lars Konieczny. 2000. Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29(6).

E. Lau, C. Stroud, S. Plesch, and C. Phillips. 2006. The role of structural prediction in rapid syntactic analysis. *Brain and Language*, 98(1).

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert Macintyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure,. In *Proceedings of ARPA Human Language Technology Workshop*.

Kentaro Nakatani and Edward Gibson. 2008. Distinguishing theories of syntactic expectation cost in sentence comprehension: Evidence from japanese. *Linguistics*, 46(1).

Joakim Nivre. 2006. *Inductive dependency parsing*, volume 34. Springer Verlag.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedinngs of HLT-NAACL*.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of EMNLP*.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational linguistics*, 27(2).

A. Staub and C. Clifton. 2006. Syntactic prediction in language comprehension: Evidence from either... or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(2).

A. Staub, C. Clifton, and L. Frazier. 2006. Heavy np shift is the parsers last resort: Evidence from eye movements. *Journal of memory and language*, 54(3).

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

A. Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*.

P. Sturt, F. Keller, and A. Dubey. 2010. Syntactic priming in comprehension: Parallelism effects with and without coordination. *Journal of Memory and Language*, 62(4).

Shravan Vasishth and Richard Lewis. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Linguistic Society of America*, 82(4).