Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities

Anonymous EMNLP submission

Abstract

Humans exhibit garden path effects: When people read temporarily structurally ambiguous sentences, they slow down when the structure is disambiguated in favor of the less preferred alternative. One prominent explanation of this, surprisal theory (Hale, 2001; Levy, 2008), proposes that slow downs like those in garden path sentences is due to the (un)predictability of each word in context. van Schijndel and Linzen (2021) find that estimates of the cost of word predictability derived from LSTM language models severely underestimate the magnitude of garden path effects. In this work, we consider whether this underestimation is due to the underweighting of the structural factors in language models' estimates of word predictability relative to humans. We propose a method for estimating syntactic predictability from a language model, allowing us to weigh the cost of lexical and syntactic predictability independently. We find that treating syntactic predictability independently from word predictability results in larger estimates of garden path effects than models with only word predictability. However, even with independently weighted syntactic predictability, surprisal-based models still greatly underestimate the magnitude of garden path effects seen in humans, suggesting that there are factors other than predictability at play in the processing cost of garden path sentences.

1 Introduction

Readers exhibit *garden path effects*: When shown a temporarily syntactically ambiguous sentence, readers tend to slow down when the sentence is disambiguated in favor of the less preferred parse. For example, a participant who reads the sentence fragment

(1) The suspect sent the file . . .

a. . . . to the lawyer.

b. . . . deserved further. investigation

can construct a partial parse in at least two distinct ways: in one reading, the verb *sent* acts as the main verb of the sentence, and the rest of the sentence may be an additional argument to *sent* (as in 1a). However, another, less likely reading has *sent the file* acting as a modifier in a complex subject that still requires an additional verb phrase to form a complete sentence (e.g.., in 1b). Prior work has demonstrated that regions such as *deserved further investigation*, which disambiguate these temporarily ambiguous sentences in favor of the modifier parse, are read slower than that same words would be in an unambiguous version of sentence such as

(2) The suspect *who was* sent the file deserved further investigation.

where the presence of *who was* signals to a reader that *sent the file* acts as a modifier (Frazier and Fodor, 1978).

One account of this phenomenon, surprisal theory (Hale, 2001; Levy, 2008), suggests that readers maintain a probabilistic representation of all possible parses of the input as they process the sentence incrementally. Processing difficulty is then the cost associated with updating this representation, which is directly proportional to the negative log probability, or surprisal, of the newly observed material under the reader's model of upcoming words. This model predicts that the slowdown associated with garden path sentences can be entirely captured by the differences in surprisal between the disambiguating region in ambiguous garden path sentences and its counterpart in an unambiguous sentence.

van Schijndel and Linzen (2021) test this hypothesis directly, estimating the surprisals associated with garden path sentences using LSTM language models (LMs) trained over large natural language

corpora. They then leverage the claim that predictability should be proportional to processing difficulty for all sentences to estimate a conversion factor between surprisals and reading times over non-garden path sentences. Using the results of this conversion, van Schijndel and Linzen (2021) find that surprisal theory, paired with the surprisals estimated by their models, severely underestimates the magnitude of the garden path effect for three garden path constructions. Moreover, the predicted reading times do not correctly predict the differences in the magnitudes of the average garden path effect for each construction, suggesting that any single set of conversion factors between surprisal and reading times would be unable to correctly predict the magnitude of the garden path effect in all three constructions. From this result, we can draw one of two conclusions: either (1) a strong version of surprisal theory cannot account for garden path effects, or (2) the estimates of predictability derived from LSTM LMs do not capture some relevant properties of garden path sentences.

In this work, we will investigate the latter possibility. In particular, we seek to determine whether the gap between the magnitude of garden path effects in humans and the magnitude that surprisal theory predicts from LM predictability estimates is due to a mismatch between how humans and LMs weigh two contributors to word-level surprisal: syntactic and lexical predictability. That is, the task of predicting the next word in large natural corpora may underestimate the importance of predicting upcoming syntactic structure to human readers (i.e., weigh lexical factors on predictability much more heavily than syntactic factors). In this scenario, since garden paths are the product of unpredictable syntactic structure rather than unpredictable lexical items, using a model's predictability estimate for the next word would lead to the underestimation of garden path effects that we see. If this is the case, this gap can be bridged by isolating the surprisal associated with the syntactic structure implied by the next word from word-level predictability, as shown in figure 1, and weighing the two factors independently of our model's language modeling objective. This follows prior work on syntactic (or unlexicalized) surprisal primarily in the context of symbolic parsers, where the probability of a structure and particular lexical item can be easily disentangled (Demberg and Keller, 2008; Roark et al., 2009). While past work has demonstrated that that

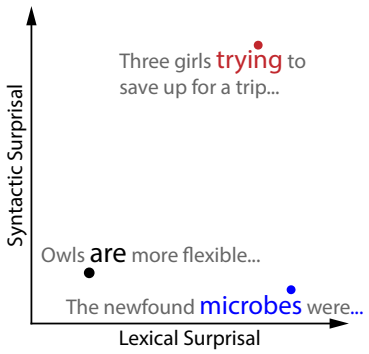


Figure 1: A depiction of the relationship between syntactic and lexical surprisal. Some words (*are*) are highly predictable in all respects. Others are unpredictable due to the syntactic structures they suggest (*trying*), and should earn a high syntactic and lexical surprisal. Words like *microbes*, on the other hand, appear in predictable syntactic environments, but are unpredictable due to their infrequency. Words like this should earn low syntactic surprisals despite having a high lexical surprisal. Since words that appear in unpredictable syntactic environments are necessarily unpredictable themselves, no words have high syntactic surprisal and low lexical surprisal.

unlexicalized suprisal from symbolic parsers correlates with measures of human processing difficulty (Demberg and Keller, 2008), simple recurrent neural networks trained to predict sequences of part-of-speech tags have been shown to track processing difficulty even more strongly (Frank, 2009).

We extend this line of work by considering LMs that are additionally trained to estimate the likelihood of the next word's supertag under the Combinatory Categorial Grammar (CCG) framework. Such supertags can be viewed as enriched part-of-speech tags that encode syntactic information about how a particular word can be combined with its local environment. We then define syntactic surprisal in terms of the likelihood of the next word's CCG supertag, and propose a method of estimating that likelihood using our modified LMs. We validate our formulation of syntactic surprisal by demonstrating that it captures syntactic processing difficulty in garden path sentences while, crucially, not tracking unpredictability that is due to low frequency lexical items. Following van Schijndel and Linzen (2021), we then use the syntactic and lexical surprisal values derived from those models to predict reading times for three types of garden path sentences. We find that the addition of our estimates of syntactic surprisal as a separate pre-

dictor does not substantially improve the ability of surprisal theory to account for garden path effects. Finally, we discuss the implications of this finding on surprisal theory and single-stage models of syntactic processing.

2 Computing Syntactic Surprisal

Our goal is to evaluate whether a measure of *syntactic* surprisal can help account for the magnitude of garden path effects. In order to do so, we will need a way to estimate syntactic surprisal in a manner that both captures our intuitive notions — an incremental representation of the predictability of the sentence's syntactic structure suggested by the next word — while being relatively simple to estimate. For our purposes, we will define syntactic surprisal in terms of the predictability of the next word's supertag under the Combinatory Categorial Grammar (CCG) formalism (Steedman, 1987). That is, the syntactic surprisal of a word w_n is

$$\text{surp}_{\text{syn}} = -\log(P(c_n \mid w_1, ..., w_{n-1})), \quad (1)$$

where c_n is the supertag of the n-th word under the CCG formalism. A CCG supertag encodes how a particular token combines with adjacent constituents in that token's sentence's syntactic structure. For example, a token with the tag S\NP combines with a constituent with the tag NP to its left to form a constituent with the tag S. Similarly, a token with the tag (S\NP)/NP combines with a constituent with the tag NP to its right to form a constituent with the tag S\NP. Since a full supertagging of a sentence often allows only one valid parse, the task of predicting a sentence's supertags has been described as "almost parsing" (Bangalore and Joshi, 1999). Thus we see incremental CCG supertagging as almost *incremental* parsing, and the surprisal of each word's supertag under a probabilistic incremental supertag predictor as a reasonable proxy for syntactic surprisal. We compare this syntactic surprisal measure with the standard token surprisal measure, which we refer to as *lexical* surprisal:

$$\text{surp}_{\text{lex}} = -\log(P(w_n \mid w_1, ..., w_{n-1})). \quad (2)$$

Note that what we call lexical surprisal captures all factors that contribute to a token's predictability, including syntactic ones: Since certain tokens can only appear in specific syntactic contexts, the probability that the next word is in one of those contexts affects the predictability of any particular word.

In order to compute syntactic and lexical surprisals for our materials, we need models that predict, given a left context, both the next token (as a standard LM does) and the next token's supertag (to compute syntactic surprisal). To do this, we train a model with both a language modeling and CCG supertagging objective and estimate the distribution over the next word's tag by marginalizing over the identity of the next word itself. Formally, for a sequence of words $w_1, ..., w_n \in W$ with supertags $c_1, ..., c_n \in C$, our model estimates the probability of the next word given all observed words, $p_{w_{n+1}} = P(w_{n+1} \mid w_1, ..., w_n)$, and the probability of the most recent word's supertag given all currently observed words, $p_{c_n|w_n} = P(c_n \mid w_1, ..., w_n)$. We then infer the distribution over the next word's supertag as

$$P(c_{n+1}|w_1, ..., w_n) = \sum_{w^*_{n+1} \in W} p_{c_{n+1}|w^*_{n+1}} p_{w^*_{n+1}} \quad (3)$$

If we knew the supertag of the next word c_{n+1}, we can then compute the surprisal of that supertag by computing $-\log P(c_{n+1} \mid w_1, ..., w_n)$. Unlike in the case of lexical surprisal, however, a word's supertag is often ambiguous during incremental processing. Consider the verb *gathered* in the following examples:

(3) The squirrels gathered near the tree.

(4) The squirrels gathered a few acorns.

In (3), we would want to assign *gathered* the supertag S\NP, indicating that *gathered* is used in its intransitive sense (a group of squirrels came together as a group) and takes no direct object. In (4), on the other hand, we would want to assign the supertag (S \NP)/NP to indicate that in this sentence *gathered* is used in its transitive sense and takes the noun phrase *a few acorns* as a direct object. When processing this sentence incrementally, a reader must maintain this uncertainty over the appropriate supertag for a word past the point at which they've learned the word's identity. As a result, a measure of syntactic surprisal that aims to model processing difficulty at a particular word should similarly take into account uncertainty over the supertag of a word even after the word itself has been processed.

Since our models are trained as supertaggers in addition to being LMs, we use the models' supertagging distribution ($p_{c_n|w_n}$) to estimate the un-

certainty over a word's tag once it has been observed. We thus compute surprisal after marginalizing the probability of the next word's supertag we computed in (3) over our uncertainty over that next word's supertag:

$$p_{c_{n+1}|w_n} = P(c_{n+1} \mid w_1, ..., w_n) \tag{4}$$

$$surp_{syn} = -\log \sum_{c^*_{n+1} \in C} p_{c^*_{n+1}|w_n} p_{c_{n+1}|w_{n+1}} \tag{5}$$

2.1 Model Architecture and Training

We train four LSTM LMs, differing only in their random seed, on both a language modeling and CCG supertagging objective.

Following Gulordava et al. (2018), the encoder is a two-layer LSTM with 650 units per layer. Each decoder consists of a single linear layer and a softmax classifier. Models were trained on supertagging using CCGBank (Hockenmaier and Steedman, 2007), a set of CCG annotations for the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993), and on language modeling over a concatenation of the Wall Street Journal portion of the Penn Treebank and the 80 million words of Wikipedia used in Gulordava et al. (2018). Language modeling and supertagging losses were weighted equally during training.

Models achieved language modeling perplexities ranging from 74.76 to 75.70 on the Gulordava et al. (2018) test set, and assigned the highest likelihood to the correct CCG supertag in the CCGBank test set between 84.1% and 84.5% of the time.

2.2 Experimental data

We evaluate our model over a subset of the Syntactic Ambiguity Processing (SAP) Benchmark (Huang et al., 2022), a dataset containing self-paced reading data from 2000 native English speakers over a variety of constructions as well as a number of filler sentences. The large size of the dataset allows us to get precise estimates of the magnitude of the garden path effect for each of the three types of garden path sentences it contains. We describe each of the three garden path constructions below.

(5) The suspect sent the file **deserved** further investigation given the new evidence.

(6) The suspect who was sent the file **deserved** further investigation given the new evidence.

Main Verb/Reduced Relative (MVRR): In (5), before reading the word *deserved*, the reader can interpret *sent the file* either as a main verb and direct object (where the subject has sent the file) or as a reduced relative clause (where the subject has had the file sent to them). This is disambiguated in favor of the reduced relative clause reading by the next word, *deserved*, which is the true main verb of the complete sentence. We can measure the processing difficulty incurred by this disambiguation by comparing the reading times at *deserved* in (5) with those at *deserved* in (6), where the relative clause *who was sent the file* is unreduced and thus unambiguous.

(7) The suspect showed the file **deserved** further investigation during the murder trial.

(8) The suspect showed that the file **deserved** further investigation during the murder trial.

Noun Phrase/Sentence (NPS): In (7), before reading *deserved*, *the file* can be interpreted as just a noun phrase acting as a direct object (where the suspect is presenting a file to someone) or as the beginning of a sentential complement (where the suspect is making a point). Again, *deserved* disambiguates this in favor of the less frequent sentential complement reading, and its counterpart in (8) avoids the ambiguity altogether by using the explicit complementizer *that* before *the file*, giving us a way to measure the slowdown associated with disambiguation.

(9) Because the suspect changed the file **deserved** further investigation during the jury discussions.

(10) Because the suspect changed, the file **deserved** further investigation during the jury discussions.

Noun Phrase/Zero (NPZ): Finally, in (9), before reading *deserved*, *changed* can be interpreted as a transitive verb taking *the file* as a noun phrase direct object (where the file was changed by the suspect), or as an intransitive verb with no direct object and *the file* as the subject of a separate clause (where the suspect was changed). *deserved* disambiguates this in favor of the less frequent intransitive reading, and introducing a comma between the clauses in the sentence's counterpart in (10) removes the ambiguity.

3 Validating Syntactic Surprisal

As our method of computing syntactic surprisal is novel, we first validate that it successfully isolates syntactic predictability from word predictability. For this to be the case, we will require that two things be true: that syntactic surprisal captures processing difficulty that is the result of syntactic unpredictability, and that syntactic surprisal is **not** redundant with lexical predictability. We will evaluate each of these desiderata in turn.

3.1 Syntactic Surprisal Captures Syntactic Processing Difficulty:

To verify that syntactic surprisal can capture syntactic unpredictability, we investigate differences in syntactic surprisal between the ambiguous and unambiguous garden path sentences in Huang et al. (2022). Since garden path effects are the result of ambiguity about the syntactic structure of a sentence, a difference in surprisal at the point of disambiguation indicates sensitivity to differences in syntactic predictability. We find these differences in all types of garden paths for lexical surprisal (as has been previously shown; Hale (2001); van Schijndel and Linzen (2021)) as well as for syntactic surprisal (Figure 2). We do not find differences in the same direction before the point of disambiguation, indicating that the differences we observe post-disambiguation are not simply the result of prior surprisal differences spilling over.

3.2 Syntactic Surprisal Captures Only Syntactic Predictability

To verify that syntactic surprisal successfully isolates syntactic factors on predictability, we make two comparisons: first to lexical surprisal, to verify that syntactic surprisal does not capture all of the variance lexical surprisal does, and second to unigram frequency, to verify that syntactic surprisal isn't driven by the frequency of specific lexical items.

Syntactical Surprisal does not capture all of lexical surprisal's variance: If syntactic surprisal captures a strict subset of the variance captured by lexical surprisal, we expect to see a subset of words with high lexical surprisal and low syntactic surprisal (in addition to words with highly correlated syntactic and lexical surprisals). This subset should represent words that are unpredictable for reasons independent of the syntactic structures they imply, whereas words that introduce infrequent syntactic structures should have high syntactic and lexical surprisals, as the unpredictability of the syntactic structure means that a word that implies that structure is necessarily unpredictable. This matches what we see in Figure 3a: the relatively frequent verb *trying* introducing a reduced relative clause has high syntactic and lexical surprisal, while infrequent nouns like *microbe* have low syntactic surprisal but high lexical surprisal.

Unigram (in)frequency does not drive syntactic surprisal: In Figure 3b, where we plot syntactic surprisals for words in filler items with their log-frequency within the Corpus of Contemporary American English (COCA; Davies (2008–)). We find a significant but small positive correlation between the two ($r = 0.064$, $t = 3.18$, $p < 0.005$), indicating that more frequent words have a *higher* syntactic surprisal — the opposite of what we'd expect if lexical (in)frequency were driving syntactic surprisal effects. This is likely due to the fact that function words, which are generally high-frequency due to their closed-class nature, typically introduce additional syntactic structure and thus have higher-than-average syntactic surprisal.

These three results — that syntactic surprisal captures garden path effects, that we find a subset of words with low syntactic surprisal and high lexical surprisal, and that we find no evidence of lexical (in)frequency driving syntactic surprisal — suggest that syntactic surprisal captures only the syntactic contributions to a word's unpredictability. We will now use syntactic surprisal in concert with lexical surprisal to directly predict the magnitude of garden path effects.

4 Evaluating Against Human Reading Times

Recall that surprisal theory assumes a linear relationship between surprisal and measures of processing difficulty such as reading times. We follow van Schijndel and Linzen (2021) and estimate a mapping between our surprisal measures and reading times by fitting linear mixed effects models to the filler (i.e., non-garden path) materials in Huang et al. (2022). In order to compare syntactic and lexical surprisal, we fit four models: one with syntactic surprisal as a predictor, one with lexical surprisal as a predictor, one with both types of surprisal, and one that ignores surprisal entirely. All four models include non-surprisal predictors — unigram frequency, word position, and word length — which

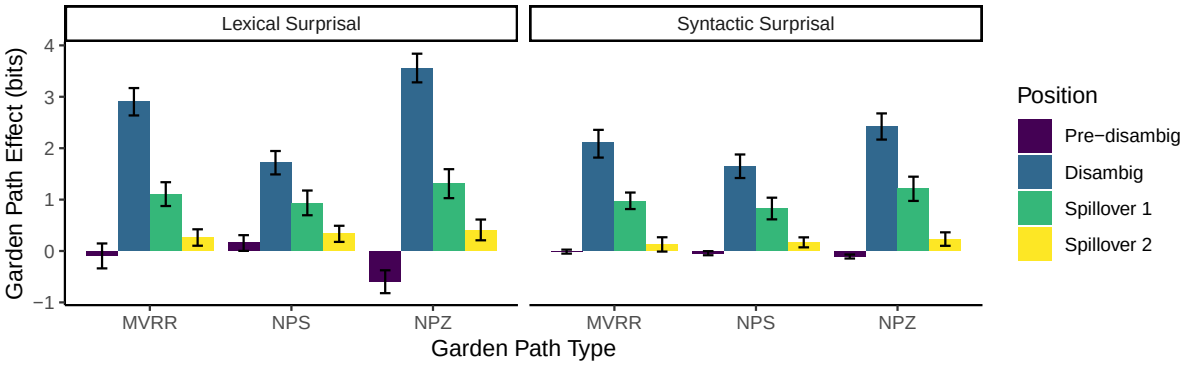
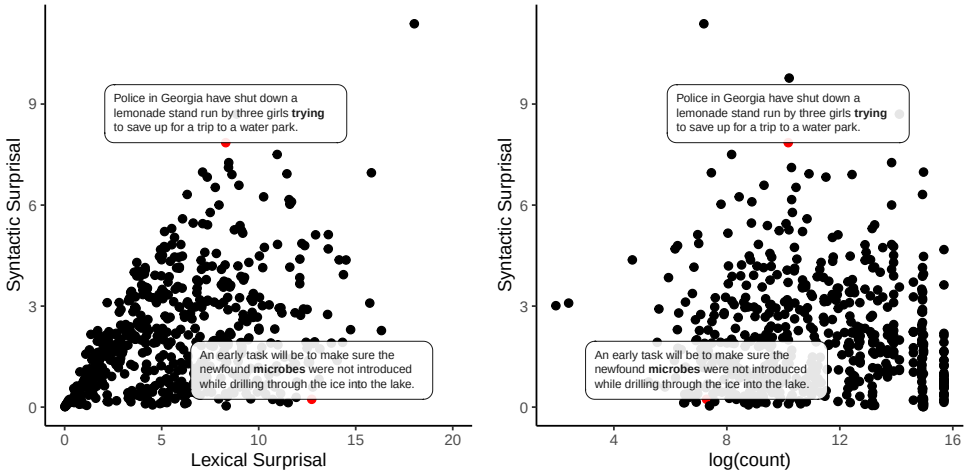


Figure 2: Differences in surprisal estimates between ambiguous and unambiguous garden path sentences at and around the disambiguating verb. Bars indicate 95% confidence intervals.



(a) Lexical vs. syntactic surprisal

(b) log unigram frequency vs. syntactic surprisal

Figure 3: Correlations between syntactic surprisal, lexical surprisal, and unigram frequency for each word in the filler items of Huang et al. (2022). Two words — one with high syntactic surprisal and high lexical surprisal and one with high lexical surprisal but low syntactic surprisal — are labeled with their context.

alone are not purported to capture garden path effects. To account for spillover effects, where processing difficulty from a word spills over to affect reading times at future words, we include all of the aforementioned factors (except word position) not only for the current word but also for the two prior words (a simplification of the technique in van Schijndel and Linzen (2021)). Further details about these models are presented in Appendix A.1. After all four of our models have been fit to the filler items, we use the estimated coefficients to predict reading times for the each of the critical items.

5 Results

Predicted RT differences from our models, as well as the RT differences observed in humans, are presented in Figure 4. Regardless of which RT conversion method is used, predicted reading time differences greatly underestimate the reading time differences observed in humans. This is unlikely to be an issue with our surprisal-to-reading-times conversion method more broadly, as at the pre-disambiguation word, RTs and predicted RTs match much more closely than in post-disambiguation regions, indicating that the difference in magnitudes is due specifically to an underestimation of the garden path effect.

Underestimation aside, if we compare amongst the various RT conversion models, we can determine whether the inclusion of syntactic surprisal in our RT-prediction models leads to a larger predicted garden path effect. To see this difference more clearly, in Figure 5 we exclude the human reading times and zoom in on the garden path effects predicted by the models. Table 1 presents the

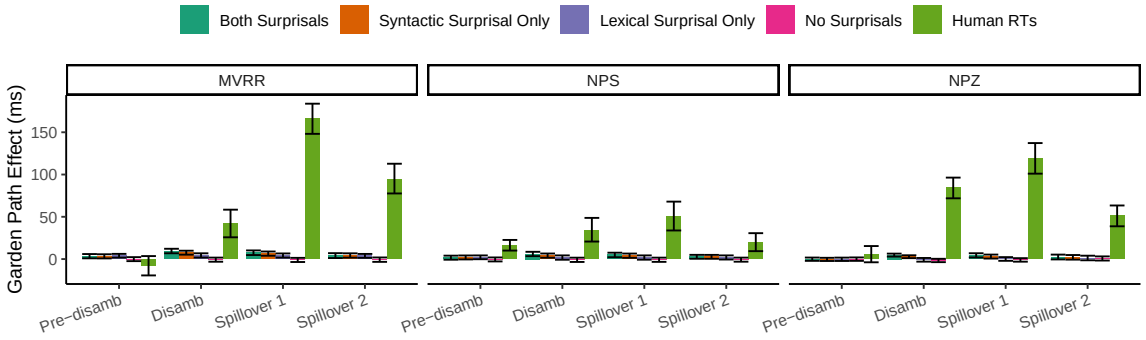


Figure 4: Human and model-predicted Garden Path Effects. Bars are the difference between the mean ambiguous and mean unambiguous reading times across participants for each condition. Error bars indicate bootstrapped 95% confidence intervals.

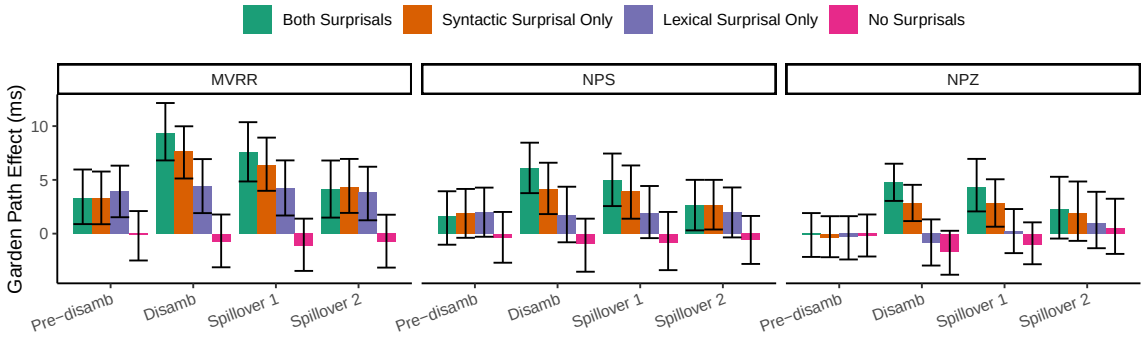


Figure 5: Predicted Garden Path Effects using each of our four RT prediction models. Bars are the difference between the mean ambiguous and mean unambiguous reading times across participants for each condition. Error bars indicate bootstrapped 95% confidence intervals.

results of a mixed effects analysis of the predicted RTs (for details see Appendix A.2). We find that (1) models containing both surprisals predicted the largest garden path effects, (2) models with only syntactic surprisal predicted slightly smaller effects, and (3) that models with only lexical surprisal or no surprisal predicted much smaller garden path effects than surprisal only models.

6 Discussion

In this paper we evaluate one explanation for the discrepancy between the magnitude of garden path effects in humans and surprisal-based estimates of those magnitudes from LSTM LMs: That, compared to human predictions, word predictability estimates from LMs underweight the importance of syntactic predictability relative to other factors. We propose a method of estimating syntactic predictability from modified LSTM LMs, validate this method's ability to match our intuitions of the measure, and compare garden path effect magnitude predictions derived from standard, lexical surprisal and syntactic surprisal. We find that while syntactic surprisal does lead to larger predicted garden path effects, model-predicted garden path effects still vastly underestimate the magnitude of garden path effects found in humans.

We define syntactic surprisal in terms of the predictability of the next word's CCG supertag — a label that indicates how that word syntactically combines with its local context. This is motivated by three desiderata: First, we want a measure that captures processing difficulty due to syntactic unpredictability. Since CCG supertags capture local syntactic structure, we hypothesize that the surprisal of that supertag is a good predictor of syntactic unpredictability. This is borne out in our evaluation of syntactic surprisal in garden path sentences, where syntactic surprisal predicts differences for our three garden path constructions. Second, since syntactic surprisal is designed to isolate syntactic predictability, we want it *not* to track purely lexical factors

	MVRR	NPS	NPZ
Both vs Syntactic Only	$\beta = 0.93, p < 0.001$	$\beta = 1.14, p < 0.001$	$\beta = 1.13, p < 0.001$
Syntactic Only vs Lexical Only	$\beta = 2.06, p < 0.001$	$\beta = 1.94, p < 0.001$	$\beta = 2.23, p < 0.001$
Syntactic Only vs Neither	$\beta = 7.09, p < 0.001$	$\beta = 4.78, p < 0.001$	$\beta = 2.95, p < 0.001$

Table 1: Differences in garden path effects over the critical region from models that predict RTs from just syntactic surprisal, just lexical surprisal, both, or neither estimated from a linear mixed effects model.

on word predictability. We find this to be the case in analyses comparing it to lexical surprisal and unigram frequency: We found (lexically) surprising, but syntactically predictable words with low syntactic surprisal, as well as a positive correlation between frequency and syntactic surprisal — the opposite of what would be predicted if syntactic surprisal was driven by unigram (in)frequency. Finally, we want syntactic surprisal to be simple to compute, which motivated us to use a measure derivable from CCG supertagging and language modeling probabilities, two simple sequence labelling tasks, rather than more sophisticated parsing objectives.

The increase in garden path magnitudes we see when using syntactic surprisal suggests that predictability estimates from LSTM LMs may indeed undervalue the importance of syntactic factors relative to humans. That is, since syntactic surprisal captures a subset of the variance that lexical surprisal does, the fact that considering syntactic surprisal when predicting human reading times leads to better performance than using only lexical surprisal suggests that the influence of syntactic factors on lexical surprisal is small relative to its ability to capture variation in human reading times. One potential explanation for this may be the difference in the tasks humans and LMs perform: While LMs need only predict words in corpora, humans must also attempt to comprehend what they read. While both tasks demand some sensitivity to syntactic structure, the need to interpret sentences may place greater importance on predicting structure, leading to a higher sensitivity to syntactic unpredictability.

While our formulation of syntactic surprisal allowed us to demonstrate the importance of a stronger emphasis on syntactic predictability, the large discrepancy between model-predicted and human garden path effect sizes indicates that there is still much work to be done if we are to explain these effects with surprisal-based theories. Future work attempting to do so can look to more explicit models of syntactic predictability than predicting CCG supertags. Architectures like the Recurrent Neural Network Grammar (Dyer et al., 2016) derive word-level predictability estimates from explicit syntactic parsing mechanisms rather than attempt to derive syntactic predictability from word-level predictability, and as a result may generate better estimates of syntactic predictability than we do.

Another possibility that suggests itself is that surprisal-based accounts of garden path effects simply cannot account for the magnitude of the slowdowns observed in humans, even if it was based on a perfect simulation of the human language model. Instead, we may need to adopt a different model of human syntactic disambiguation. One set of alternatives are *two-stage, serial* models of processing (Frazier and Fodor, 1978). In such a model, when readers first read through the ambiguous fragment of the sentence, they commit to a small set of preferred parses. When they reach a disambiguating region where all of the parses they have committed to are no longer consistent with the input, a reader would engage a separate, costly reanalysis process in order to construct a new partial parse consistent with the all of the currently available input. The processing cost associated with this reanalysis procedures incurs a slowdown in reading times that does not occur in an unambiguous sentence where the incorrect initial parse is not available, resulting the garden path effects that we observe. Unlike surprisal-based accounts, however, it is unclear how to derive quantitative predictions for the size of garden path effects from current two-stage accounts. As a result, it is difficult to know whether the quantitative mismatches between surprisal-accounts and human reading times that we observed should be taken as evidence for a two-stage account. This further highlights the need for more precise quantitative accounts of two-stage serial models we can evaluate surprisal accounts against.

References

Srinivas Bangalore and Aravind Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Mark Davies. 2008–. The Corpus of Contemporary American English (COCA).

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

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent Neural Network Grammars. In *North American Chapter of the Association for Computational Linguistics*.

Stefan Frank. 2009. Surprisal-based Comparison between a Symbolic and a Connectionist Model of Sentence Processing. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank:a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Grusha Prasad Christian Muxica, Brian Dillon, and Tal Linzen. 2022. SPR mega-benchmark shows surprisal tracks construction- but not item-level difficulty. In *35th Annual Conference on Human Sentence Processing*.

Roger Levy. 2008. Expectation-Based Syntactic Comprehension. *Cognition*.

MP Marcus, B Santorini, and MA Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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 the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Mark Steedman. 1987. Combinatory grammars and parasitic gaps. *Natural Language & Linguistic Theory*.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

A Appendix

A.1 Converting Surprisals to Reading Times

In order to gauge the impact of syntactic surprisal on the predicted reading time at word n, rt_n, we fit four mixed effects models over the filler data: one containing only lexical surprisal (s_n^{lex}), one containing only syntactic surprisal (s_n^{syn}), one containing both, and one containing neither. As reading times are sensitive to other features of the word being read like unigram frequency (f_n), position in sentence p, and length in characters (c_n), we include those variables as additional factors in the regression. In order to account for spillover effects, where processing difficulty from a word often surfaces in the reading times of subsequent words, we include all of the aforementioned factors for the prior two words. We additionally include random intercepts by item and by participant, as well as random slopes by item for all of the surprisal fixed effects. This gives us the following linear mixed effects model formulas:

$$\begin{aligned} rt_n \sim\ & f_n * c_n + f_{n-1} * c_{n-1} \\ & + f_{n-2} * c_{n-2} + p \\ & + (1 \mid \text{item}) + (1 \mid \text{participant}) \end{aligned} \qquad \text{(neither)}$$

$$\begin{aligned} rt_n \sim\ & s_n^{lex} + s_{n-1}^{lex} + s_{n-2}^{lex} \\ & + f_n * c_n + f_{n-1} * c_{n-1} \\ & + f_{n-2} * c_{n-2} + p \\ & + (1 + s_n^{lex} + s_{n-1}^{lex} + s_{n-2}^{lex} \mid \text{item}) \\ & + (1 \mid \text{participant}) \end{aligned} \qquad \text{(lexical)}$$

$$\begin{aligned} rt_n \sim\ & s_n^{syn} + s_{n-1}^{syn} + s_{n-2}^{syn} + \\ & + f_n * c_n + f_{n-1} * c_{n-1} \\ & + f_{n-2} * c_{n-2} + p \\ & + (1 + s_n^{syn} + s_{n-1}^{syn} + s_{n-2}^{syn} \mid \text{item}) \\ & + (1 \mid \text{participant}) \end{aligned} \qquad \text{(syntactic)}$$

$$
\begin{aligned}
rt_n \sim\ & s^{lex}_n + s^{lex}_{n-1} + s^{lex}_{n-2} \\
& + s^{syn}_n + s^{syn}_{n-1} + s^{syn}_{n-2} \\
& + f_n * c_n + f_{n-1} * c_{n-1} \\
& + f_{n-2} * c_{n-2} + p \qquad \text{(both)} \\
& + (1 + s^{lex}_n + s^{lex}_{n-1} + s^{lex}_{n-2} \\
& + s^{syn}_n + s^{syn}_{n-1} + s^{syn}_{n-2} \mid \text{item}) \\
& + (1 \mid \text{participant})
\end{aligned}
$$

These models were fit using filler data from Huang et al. (2022), and the coefficients from each model were used to predict reading times for all of the critical, garden path items from the corresponding surprisals, frequencies, lengths, and positions.

A.2 Statistical Analysis of Predicted RTs

To analyze the predicted reading times that come from our four models of surprisal-to-reading time conversion, we fit three separate linear mixed effects models: one over MVRR garden paths, one over NPS garden paths, and one over NPZ garden paths. Each model includes fixed effects of ambiguity and the types of surprisals used in predicting reading times: syntactic surprisal only, lexical surprisal only, both surprisals, or neither. Crucially, we include the interaction between these two factors, representing how our choice of surprisal-to-RT conversion model affects the size of the predicted garden path effect. We additionally include random intercepts by item and by participant. This results in the following mixed effects model formula:

$$
\begin{aligned}
pred\_rt \sim\ & ambiguity * model \\
& + (1 \mid \text{item}) + (1 \mid \text{participant}).
\end{aligned}
$$

Since we have four different models converting between surprisals and RTs, we estimate three contrasts for the interaction term: the model with both surprisals vs. the model with only syntactic surprisals, the model with only syntactic surprisals vs. the model with only lexical surprisals, and the model with only lexical surprisals vs. the model with neither surprisal. The estimated magnitude (represented by the β coefficient) as well as significance of the difference for each of these contrasts is reported in the main text in Table 1.