The model showed the garden paths deserved further investigation:
Converting Surprisals to Reading Times for Garden-Path Sentences

Andrew Perun, Mandy Osuji, Rishika Veeramachaneni

Abstract

We observe the garden path effect as a slowdown in reading time (RT) when ambiguous constructions are disambiguated in favor of a less likely structure. Surprisal theory (Hale, 2001; Levy, 2008) provides an explanation for this phenomenon, saying that this slowdown can be entirely captured by comparing the negative log probability, or surprisal, of the disambiguating regions of the original and disambiguated forms of a garden path (GP) sentence. Previous work has aimed to apply surprisal theory in order to predict reading times for difficult constructions merely through the act of language modeling and the use of the surprisal metric. We aim to address the matter of converting from surprisal values to RT, and investigate how the choice of linear mixed-effects model training data factors into the overall prediction of garden path effects. In doing so, we seek to answer why past methods of predicting human RT using language model (LM) surprisal fail across GP sentences. We experimented with expanding the training data pool to include GP constructions, and found that doing so only improved predictions a small amount. We see that there still remains a large gap between our predicted reading times and the actual human data.

1 Background

In the broadest of terms, language is pivotal as a conduit for understanding and being understood. Examining how humans process sentences allows us to investigate the must-haves of language, and come away knowing more about the process of interpretation. Specifically, it makes sense to look at edge cases, or cases that cause the reader to struggle, in order to clarify aspects of how processing works. We will interrogate a model of real-time human processing of written language by looking at constructed examples that result in processing difficulty.

Even when words are presented in a linear sequence, there is evidence that latent abstract structures play some part in human language processing (Chomsky, 1957). To derive well-formed sentences, an acceptable structure must be imposed. The paradigm in which we operate throughout this paper assumes that an individual makes assumptions about the structure of a given sentence in real time as each word is read. However, in certain constructions, when the nth word within a stream is received, the assumed most likely structure is no longer viable and must be abandoned. For example, say an individual reads a sentence that begins:

(1) The officer awarded the honor …

The structure that the individual would likely be assuming in their mind would be one where awarded is the main verb of the sentence, such as a sentence which continues:

(a) … to her son.

This is a reasonable abstraction to have at hand; since this is a common structure, it makes sense that the reader would be most expecting the sentence to continue in such a way. However, the sentence could also go on to say

(b) … received a standing ovation.

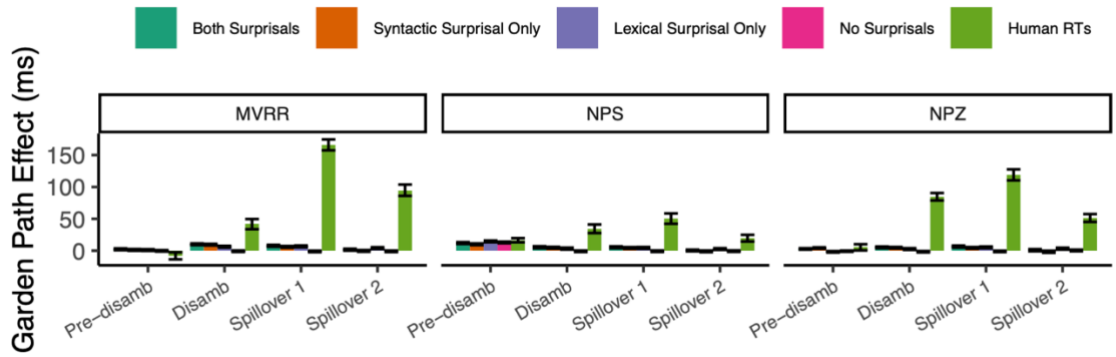


Figure 1. Arehalli et al.'s results. The green empirical RTs dwarf the MEM predicted RT.

In (1b), the appropriate parse would be one where awarded the honor modifies the subject. On the whole, the sentence is one which has an equivalent meaning to

(2) The officer who was awarded the honor received a standing ovation.

Previous work has shown that constructions such as (1b) are read more slowly than their disambiguated form, (2). Such "garden path" sentences with temporary syntactic ambiguities that are at some distinct point disambiguated in favor of an unusual (less likely) structure tend to cause processing difficulty. This difficulty is observed as longer (larger) reading times (RT) for sentences where this delayed disambiguation occurs, when compared to unambiguous alternatives of the same sentence. The garden path effect is quantified as this difference in RT between the ambiguous and unambiguous forms. This difference has been observed in human participants, and several theories of processing have been proposed to model this slowdown

.

One of the most prominent, surprisal theory (Hale, 2001; Levy, 2008), proposes a proportional relationship between processing difficulty and the negative log probability of a word given its context. It aims to quantify the cost of updating the representation structure once the disambiguating word is read. The disambiguating word is the one at which the previously assumed structure can no longer be upheld. For example, in (1b) the disambiguating word would be received. Surprisal theory claims that a metric of surprisal, defined as:

$$\text{surprisal}(w_i) = -\log P(w_i|w_1 \ldots w_i-1)$$

stands to fully explain the slowdown in RT observed for sentences which exhibit garden path effects. Because of this, past work has expected that language-model-calculated surprisal estimations would be able to fully capture garden path effect. Note that throughout, our constructions of interest are the same three types of garden-path sentences as identified in Arehalli et al. (2022). That is, Main Verb/Reduced Relative (MVRR), Noun Phrase/Sentence (NPS) and Noun Phrase/Zero (NPZ). See that paper for detailed descriptions of these forms.

2 Introduction & Motivation

Arehalli et al. (2022) considers past results (Van Schijndel and Linzen (2021)) of severe underestimation from LM-based surprisal estimations of garden path effects as explicable by one of two interpretations:

- Surprisal theory alone cannot account for garden path effects.
- Predictability estimates created from the LMs fail to fully capture some human incentives in processing garden path sentences.

Arehalli et al. explore possibility two by expanding the surprisal measurement past LM's objective of next-word prediction to mimic human's emphasis on syntactic surprisal during processing. This was implemented via giving their LM an auxiliary objective of estimating likelihood for the next word's supertag. They then explored

how the past surprisal metric, now renamed lexical surprisal, and the new syntactic surprisal metric performed when fed as input features, alongside unigram frequency, word length, and word position, to a linear mixed-effect model predicting RT.

Garden path effects were then calculated from these models' predicted RT differences across an ambiguous garden path construction of a sentence and an unambiguous construction at the same disambiguating region. Arehalli et al.'s methods and updated surprisal measurement predicted larger magnitudes of garden path effects and correspondingly slower predicted RT through the linear model. However, their model predictions still vastly underestimated observed human RT as seen in Figure 1.

One explanation for this discrepancy they offer is a return to possibility (1): the inadequacy of surprisal as a measure for capturing the magnitude of the garden path effect as seen in humans. The persistence of underestimation despite the improvements their surprisal measurement offered in the direction of a more perfect simulation of human processing and prediction could suggest this. Alternatively, there is also plenty of room to further liken the LM predictability estimated to humans by increasing the importance of other human processing cues. Before either of these conclusions are reached and pursued, however, the strength of the linear mixed-effect model step for mapping surprisal to RT deserves a more critical eye.

We hold that the discrepancy might be explained by a mismatch in how Arehalli et al. built their regressions and how they were usd. Their models were trained exclusively on filler or unambiguous data. It received no garden path constructions during training, but the testing task involved using the model on garden path sentences and their unambiguous pairs. We anticipate that given the different classes of sentences constituting the train and test sets, the underperformance of the model may actually be the expected result from an incorrect use of the model out of distribution.

Statistical analysis can allow us to investigate this as a source of the magnitude discrepancy and direct exploration of alternative models. We aim to explore why the predicted RT are so low and look to explore how the choices in mapping from surprisal values to RT influence overall prediction of garden path effects to increase magnitude.

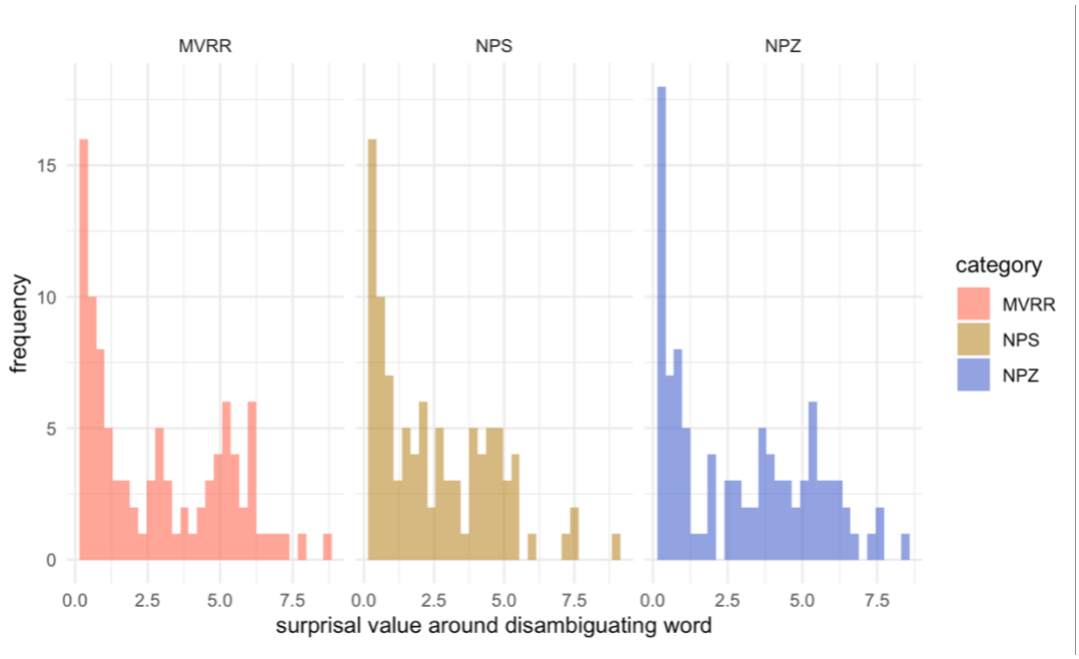
3 Methods

3.1 Features of Surprisal and Reading Time Distribution

Recall that, to predict reading times from lexical and syntactic surprisal in the disambiguating word's critical region, Arehalli et al. employed four linear mixed-effect models trained on the filler items of Huang et al. (2022). The model utilizing only surprisal most accurately predicted reading times and garden path effects, though still did not close the gap between model predictions and empirical reading times.

To explain the model's shortcomings, we employ the Kolmogorov–Smirnov test to compare the distributions between the model's training data - surprisals produced from filler, non-garden path text - and test data - surprisals produced from garden path sentences (Virtanen et al., 2020). In this test, we find that the differences in the word-specific training and test surprisal distributions are statistically significant. These distributions also have significantly different variances using the Fligner-Killeen t-test, indicating that the model was tested on out-of-sample data (Fligner and Killeen, 1976). All tests were significant at $p < 0.05$.

Figure 2. Syntactic surprisal distributions, sampled from the area of interest in ambiguous garden path constructions.



Additionally, we find the distributional and variance differences in reading times between the filler data and garden path data to be significant, both at $p < 0.05$ (Virtanen et al., 2020; Fligner and Killeen, 1976). Since Arehalli's training and test data violate the required presumption that residuals are identically and independently distributed, predictions for garden path cases are likely less accurate.

We also perform pairwise comparisons of the word-specific surprisal samples from each area of interest in the garden path case. We find no two are sampled from the same distribution nor have similar variances (Virtanen et al., 2020; Fligner and Killeen, 1976). All tests were significant at $p < 0.05$. The surprisal distributions are shown in Figure 2.

3.2 Experiment 1

To rectify the discrepancy between the empirical and syntactic model-predicted reading times, we proceed by training linear mixed-effects models on both filler and garden path data. That is, for each of the three garden path cases, we train a model on surprisals and reading times from the filler data and two garden path cases and test on the excluded case. We hypothesize that this may slightly improve accuracy in reading time predictions, as some sentences included in the training set have similar syntactic constructions to-though not necessarily the same ambiguity characteristics of - test set data.

3.3 Experiment 2

We train three models using surprisals and reading times from two garden path cases, completely excluding filler data. Each model is then tested on the unseen garden path case. As a control, we also evaluate the models' ability to predict garden path effects for in-sample data - namely, the garden path constructions that were included in the training data.

4 Results

Our linear mixed-effect models improved upon the estimate found in Arehalli et al. (2022), but we found that a large gap still remains between RT predictions by our model and empirical human data, shown in Figure 3. From both experiments, including garden path examples in training led to increased RT predictions on ambiguous constructions, and therefore a larger estimated garden path effect. The models from Experiment 2 exhibit an average percent error of 68%, an improvement from the 90%+ percent error produced by models from Arehalli et al. In addition, if we look across the models, they weighed parameters differently depending on the type of GP constructions they had access to. We further observed that testing the models on in-sample data yields garden path effects shown in Figure 4. These models exhibit an average percent

Figure 3. Empirical and model-predicted garden path effects for the three garden path constructions, given by difference in reading time between ambiguous and unambiguous cases.

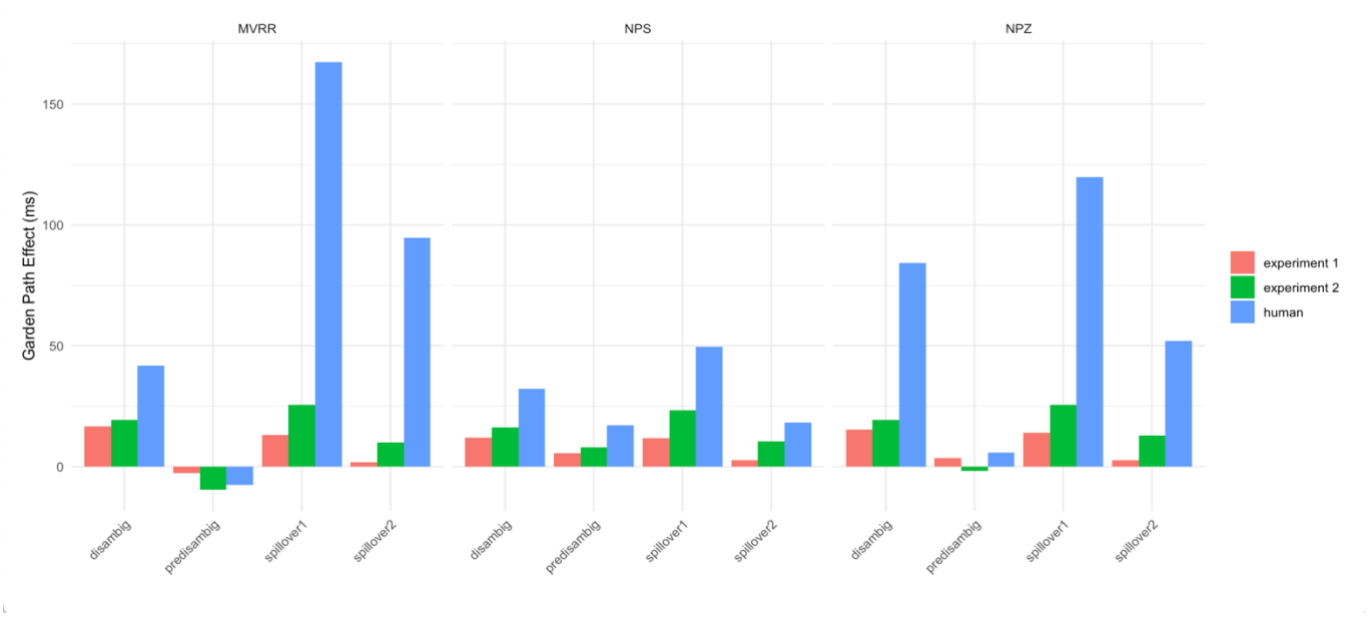
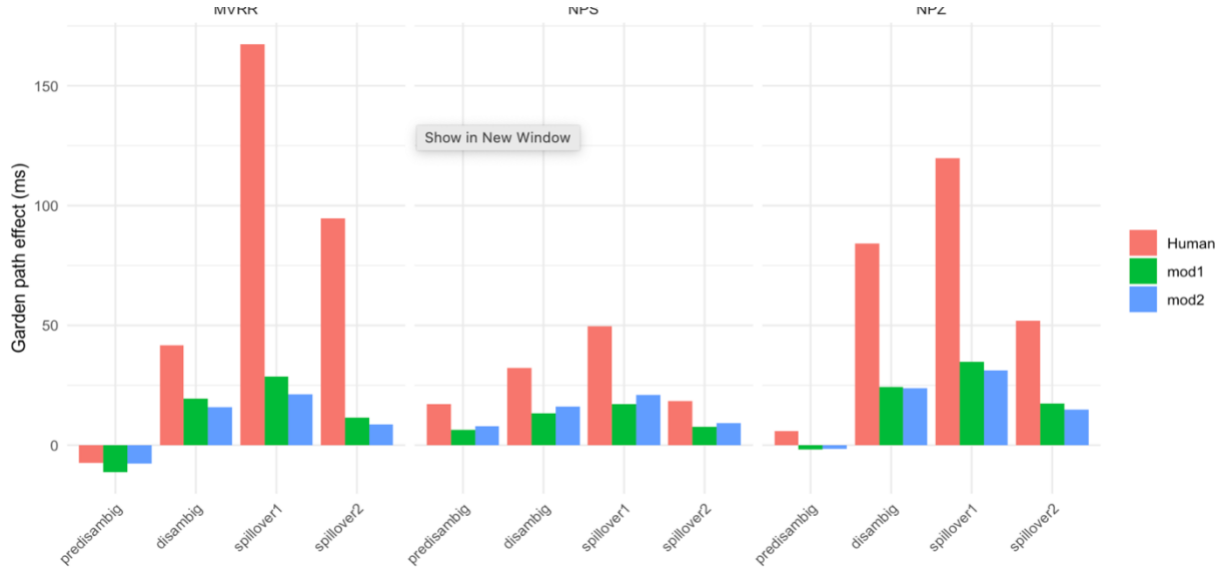


Figure 4. Predicted garden path effects for constructions included in models' training sets, with empirical effects for reference. Since a particular construction is included in two models' training sets, two predicted garden path effects are included.



error of 68.5, suggesting that though the same effect is observed regardless of garden path type, the cases are not sufficiently similar for predictions to generalize across cases when considering only syntactic surprisal.

5 Discussion

In our experiments, we consistently failed to predict reading times that led to a garden path effect of appropriate magnitude compared to human data. Because we lack improvement between models with varying training sets, the method of obtaining syntactic surprisal may be called into question. The models that produced syntactic surprisal values were trained on a CCG supertagging objective and were intended to isolate processing difficulty due to syntactic context. However, Arehalli et al. (2020) acknowledge that using a different approach to estimating syntactic surprisal could be worthwhile. Secondly, Arehalli et al. claim that syntactic surprisal captures processing difficulty by

observing its distributions in the 4-word area of interest in garden path sentences, but provide little evidence for why it necessarily captures syntactic processing difficulty over filler cases. It may be true that supertag-trained models exhibit unwanted behavior over these larger datasets, despite estimating surprisal predictably on the ambiguous cases.

Arehalli et al. also found that, around the disambiguating region, syntactic surprisal follows a similar rise-fall pattern for the three garden path types. However, upon statistical testing for distributional equivalency, we find that they are pairwise incomparable in the disambiguating region both when tested altogether and when separated by word position. Training models using multiple garden path cases then seems unjustifiable, as further demonstrated in the ineffectiveness of models in Experiment 2.

On the contrary, regardless of the garden path case, surprisal distributions are sufficiently distinct from non-garden path cases that training models on datasets nonselective towards garden path construction may still yield respectable results - that is, models that either are confined to individual garden path constructions or that don't distinguish between cases during training. A potential problem with the first approach is data unavailability. In the SAP dataset, for each garden path case, choosing specific garden path sentences to form restricts the sentence pool substantially, since their syntactic structure is quite specific. In the second approach, a model trained on all three types of garden path cases could perhaps raise RT prediction accuracy further.

Since the relationship between surprisal and RT is not preserved between filler sentences and garden path cases, nor between types of garden path cases, utilizing surprisal may not be the best approach for out-of-sample test sets in general. This is because models in Experiment 2 still did not produce garden path effects comparable to those of human values. We take these results as evidence that syntactic surprisal has not yet exhibited empirical justification in a predictive application across tested datasets using linear mixed-effects models.

This raises the question of the appropriateness of surprisal theory in general. Though it attractively and neatly explains processing difficulty by quantifying the abstract cost associated with updated syntactic representations while reading, the results from LM-based applications of surprisal theory give us pause. However, the failure of our methods to apply the theory satisfactorily led us to conclude other processing theories may be worth exploring. For example, one explanation for the garden path effect stems from the assumption that the reader may initiate a separate nonlinear reanalysis process at disambiguating phrases (Frazier and Fodor, 1978). If a process like this could be quantified with an enhanced dataset, including reading times and metrics capturing non-linear processing while reading, a trained model could more accurately estimate the true garden path effect. It may be the case that the baseline assumptions made by surprisal theory do not hold when words are presented independently and successively; a comprehensive account of humans' navigation of syntactic ambiguity is necessary. Regardless, more work remains to close the gap between the predicted and empirical garden path effect.

6 References

Arehalli, Suhas, Brian Dillon, and Tal Linzen. "Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities." arXiv preprint arXiv:2210.12187, 2022.

Chomsky, Noam. Syntactic structures. Mouton, 1957.

Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325.

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, Santa Cruz, California. Society for Human Sentence Processing.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, et al. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3):261–272.

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.

Michael A. Fligner and Timothy J. Killeen. 1976. Distribution-free two-sample tests for scale. Journal of the American Statistical Association, 71(353):210–213.