REGULAR ARTICLE - PREPUBLICATION DRAFT

Journal Section

Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty

Marten van Schijndel PhD[1] | Tal Linzen PhD[2]

[1]Department of Linguistics, Cornell University, Ithaca, NY, 14853, USA

[2]Department of Linguistics and Center for Data Science, New York University, New York, NY, 10003, USA

Correspondence

Marten van Schijndel PhD, Department of Linguistics, Cornell University, Ithaca, NY, 14853, USA

Email: mv443@cornell.edu

Funding information

The disambiguation of a syntactically ambiguous sentence in favor of dispreferred parse can lead to slower reading at the disambiguation point. This phenomenon, referred to as a garden path effect, has motivated models in which readers only maintain a subset of the possible parses of the sentence; reverting to a discarded parse requires costly reanalysis. More recently, it has been proposed that the garden path effect can be reduced to surprisal arising in a fully parallel parser: words consistent with the initially dispreferred but ultimately correct parse are simply less predictable than those consistent with the incorrect parse. The surprisal account is more parsimonious since predictability has pervasive effects in reading far beyond garden path sentences. Crucially, this account predicts a linear effect of surprisal: the difficulty experienced by readers should be proportional to the difference in word surprisal between the ultimately correct and ultimately incorrect interpretations. To test this prediction, we estimated word-by-word surprisal using recurrent neural network language models, comparing those estimates to self-paced reading times for three garden path constructions. While surprisal successfully predicted the existence of garden path responses, it severely underpredicted the magnitude of all of the garden path effects. Further, the relative size of the predicted effects was inconsistent with the relative size of the responses in humans, indicating that a differently scaled linking function would not be able to predict the response magnitudes either. These results support two-stage processing models in which recovery mechanisms beyond predictability are involved in processing garden path sentences.

KEYWORDS

self-paced reading, garden path, neural networks

1 | INTRODUCTION

Language is rife with ambiguity. In many cases, when the beginning of a sentence is compatible with multiple syntactic parses, readers consistently prefer one of those parses to the alternatives. Consider the following sentence:

(1) Even though the girl phoned the instructor was very upset with her for missing a lesson.

Readers tend to initially prefer the interpretation of (1) in which the girl phoned the instructor; in other words, they parse *the instructor* as the direct object of *phoned*. But when they reach the subsequent verb *was*, it becomes clear that this initially preferred interpretation leaves no viable subject for this verb. This leads to elevated reading times at the disambiguating region *was very upset*, compared to the same words when encountered in the following, minimally different, unambiguous sentence:

(2) Even though the girl phoned, the instructor was very upset with her for missing a lesson.

In example (2), the comma forces an intransitive interpretation of *phoned*: readers are very unlikely to consider the interpretation in which the girl phoned the instructor. Following earlier work, we will refer to the words *was very upset* as "the critical region," and to the difference in reading times on this region between (1) and (2) as a *garden path effect* (Bever, 1970).

Garden path effects have motivated cognitive theories in which, at each point of in the sentence, readers only consider one of the possible partial parses of the sentence (Frazier and Fodor, 1978; Pritchett, 1988), or only consider a subset of those parses (Gibson, 1991; Jurafsky, 1996). In those theories, processing difficulty is a consequence of the reanalysis required to reconstruct the parse that was discarded, or not considered in the first place, but later turned out to be the correct one (Pritchett, 1988; Gorrell, 1995; Sturt and Crocker, 1996; Sturt, 1997; Bader, 1998). We refer to these theories as *two-stage accounts*.

More recently, accounts such as surprisal theory (Hale, 2001; Levy, 2008) and the entropy reduction hypothesis (Hale, 2003) have attempted to derive garden-path effects from a single unified mechanism, typically based on a fully parallel probabilistic parser. Under such *one-stage accounts*, readers do not discard dispreferred parses; rather, they maintain those parses, but associate them with a lower probability compared to that of the preferred parse. Processing difficulty on every word in the sentence, including the disambiguating words in garden path sentences, arises from the extent to which the word shifts the reader's subjective probability distribution over possible parses: "the same sorts of phenomena treated in reanalysis and bounded parallelism parsing theories fall out as cases of the present, total parallelism theory" (Hale 2001, p. 6, referring to surprisal theory). If such a one-stage theory is consistent with the empirical data, it is arguably preferable to two-stage models on parsimony grounds: a theory based on a single mechanism is simpler than a theory based on two mechanisms. In the case of surprisal, for example, if word predictability—an independently motivated predictor of reading times—can account for reading behavior in garden path sentences, there is no reason to posit an additional reanalysis mechanism that comes into play only at the point where temporarily ambiguous sentences are disambiguated.

The goal of the present article is to investigate the viability of one-stage accounts of garden path effects. We focus in particular on surprisal, which has had greater success accounting for disambiguation difficulty than entropy reduction (Linzen and Jaeger, 2016). Under the surprisal account, syntactic disambiguation difficulty arises as a special case of the pervasive effects of word predictability in language comprehension (Ehrlich and Rayner, 1981; Demberg and Keller, 2008; Roark et al., 2009; van Schijndel et al., 2014). Applied to the comparison between the ambiguous sentence (1) and its unambiguous counterpart (2), for example, surprisal theory posits that the word *was* is read more slowly in (1) simply because it is less predictable in that context, as it is only consistent with a low-probability parse (Hale, 2001; Levy, 2013). Previous work has demonstrated that the words of the critical region are indeed less predictable in temporarily ambiguous sentences than in unambiguous controls. For example, Levy (2013) showed this to be true for the so-called NP/Z ambiguity illustrated in example (1) above, and concluded that "surprisal theory correctly predicts the difference in processing difficulty due to... garden pathing" (Levy, 2013, p. 94).

We argue that the conclusion that garden path effects can be *reduced* to predictability is premature, for two reasons. First, if surprisal is the only factor that explains disambiguation difficulty in garden path sentences, word predictability would need to account for the differences in difficulty across different types of temporarily ambiguous sentences. For example, just like in (1) above, in the following sentence a noun phrase (here *the contract*) is initially more likely to be interpreted as a direct object, but after the disambiguating word (*would*) this noun phrase needs to be reanalyzed as the subject of a subordinate clause:

(3) The employees understood the contract would be changed very soon to accommodate all parties.

Despite the superficial similarity of sentences such as (3), referred to as NP/S sentences, to NP/Z sentences such as (1), disambiguation causes significantly less processing difficulty in NP/S than NP/Z sentences (Pritchett, 1988; Sturt et al., 1999). Two-stage models have attributed this difference to properties of the second-stage reanalysis mechanism: certain syntactic restructuring operations are argued to be more costly than others (Pritchett, 1988; Bader, 1998). This option is not available to one-stage accounts: the surprisal hypothesis can only derive the greater disambiguation difficulty observed in NP/Z sentences if the difference in the predictability of *was* between the ambiguous NP/Z sentence and its unambiguous control is greater than the analogous difference in NP/S sentences.

A second challenge that faces one-stage models is the need to account for the full *magnitude* of the garden path effect observed in each type of ambiguity. The surprisal hypothesis predicts that the same (logarithmic) relationship should hold between predictability and reading time, regardless of whether the sentence is ambiguous or unambiguous: all else being equal, halving the conditional probability of a word in context (from p to $p/2$) should cause reading times to increase by a constant increment, regardless of the word's syntactic role and its conditional probability p. Smith and Levy (2013) report that dividing predictability by two—an additional "bit" of surprisal—leads to a slowdown of approximately 4 ms in self-paced reading experiments. By contrast, the garden path effects reported in the literature are often on the order of magnitude of dozens of milliseconds; for example, Grodner et al. (2003) report a 70 ms garden path effect for NP/Z sentences. For surprisal to explain such a difference, the surprisal of *was* in (1) needs to be about 70 ms / 4 = 17.5 bits higher than the surprisal of the same word in (2). Given the logarithmic relationship between surprisal and conditional probability, this means that the probability of *was* needs to be $2^{17.5} \approx 185,000$ times higher in the ambiguous sentence than in the unambiguous one. It is an open question whether the difference in the predictability of the critical region between ambiguous and unambiguous sentences is in fact quite this large.

To use the typology of cognitive model predictions proposed by Padó et al. (2009), we can say that these challenges to surprisal theory arise from the fact that it not only makes *qualitative* predictions—about the existence of a processing difficulty—but that it also makes *relative* predictions about the degree of processing difficulty in different contexts, and *absolute-quantitative* predictions about the precise magnitude of that processing difficulty. This is clearly a virtue of surprisal theory compared to many verbal models, which are much harder to falsify. In this work, we investigate surprisal's ability to predict, in each of these three senses, the garden path effects that are observed in self-paced reading studies.

1.1 | Estimating predictability using computational language models

How can we obtain quantitative estimates of the predictability of a word? Traditionally, predictability estimates were obtained by asking participants to perform a cloze task (Taylor, 1953). To estimate the predictability of *was* in (1), for example, participants would be asked to complete the fragment *Even though the girl phoned the instructor*. The probability of *was* in context would then be estimated as equal to the proportion of participants who completed the fragment with *was* compared to those who used a different completion. While this method is effective for distinguishing highly predictable words (e.g., $\mathrm{P}(w|\mathrm{context} = 0.8)$) from moderately predictable words (e.g., $\mathrm{P}(w|\mathrm{context} = 0.1)$), it is not effective for making distinctions among lower probability words, such as the disambiguating words in different types of garden path sentences: even if we assume, contra certain serial parsing theories (Frazier, 1979), that participants in the cloze experiment occasionally consider the dispreferred parse, millions of participants may be required to accurately estimate the very low probabilities that most likely characterize the disambiguating words in garden path sentences.

An alternative approach to estimating the predictability of words relies on probabilistic *language models*, computational models that use a large training corpus to define probability distributions over sequences of words (Goodman, 2001). Such models are better positioned than cloze tasks to estimate continuation probabilities on the order of magnitude of 2^{-18} (as may be needed to derive 70 ms effect size).

Probabilistic language models can be based on a range of different computational architectures. Many of the words in typical sentences can be predicted well from local context using n-gram models, which are based on counting short word sequences in a corpus (Goodman, 2001; Smith and Levy, 2013). By contrast, estimating predictability in syntactically complex sentences requires models that are sensitive to syntactic structure. Most work on syntactically complex sentences in computational psycholinguistics has relied on language models based on probabilistic grammars (Stolcke, 1995; Hale, 2001). Recently, recurrent neural network language models (RNNs; Elman, 1991; Mikolov et al., 2010) have been shown to make remarkably accurate word predictions compared to earlier classes of language models (Jozefowicz et al., 2016). While such models are not designed or trained with explicit syntactic annotations, recent empirical studies have shown that these models are sensitive to a number of structural properties of the sentence (Linzen et al., 2016; Gulordava et al., 2018; Wilcox et al., 2018; Futrell et al., 2019). Such highly accurate language models open up the possibility of deriving more precise predictability estimates for garden path sentences than was possible with earlier grammar-based language models.

1.2 | Overview of experiments

To test the surprisal account of garden path effects, we use surprisal estimates derived from RNN language models (described in Section 2.3) to model the results of publicly available self-paced reading data (Section 2.2). The data includes reading times for NP/Z sentences, NP/S sentences, and sentences with ambiguous reduced relative clauses, modeled after the classic ambiguity *the horse raced past the barn fell* (MV/RR sentences, Bever 1970); these constructions are described in more detail in Section 2.1. To estimate the overall correlation between language model surprisal and reading times, we use reading times on filler sentences that do not contain garden path ambiguities; we then apply the same correlation coefficient to garden path sentences. In calculating the slowdown that can be attributed to a particular unpredictable word, we pay careful attention to potential spillover effects, where the processing difficulty on an earlier word affects reading times on a later word (Section 2.4).

To anticipate our results, RNN surprisal correctly predicted a slowdown in the disambiguating region of ambiguous sentences, compared to unambiguous controls, in all three constructions; in other words, when averaged over the disambiguating region, the qualitative predictions of the surprisal account of garden path effects were borne out. But the relative and absolute-quantitative predictions were not. Surprisal underestimated the magnitude of the slowdown in all three constructions, with significant variability across constructions: the discrepancy was small in NP/S, moderate in MV/RR, and very large in NP/Z. RNN surprisal predicted numerically larger disambiguation difficulty in NP/S than NP/Z sentences, the opposite pattern from humans. Finally, even at the qualitative level of explanation, surprisal did not predict the detailed word-by-word contour of the garden path effect over the disambiguating region. With important limitations that we discuss below, these results challenge the hypothesis that processing difficulty in garden path sentences can be reduced to predictability, and suggest that the disambiguation of garden path sentences may engage additional reanalysis mechanisms.

2 | METHODS

2.1 | Materials

We study three classic types of temporary syntactic ambiguities (Frazier, 1979). The first type is the NP/S ambiguity, illustrated in (4a):

(4) a. The employees understood the contract <u>would be changed</u> very soon to accommodate all parties.
b. The employees understood that the contract <u>would be changed</u> very soon to accommodate all parties.

The label NP/S reflects that fact that the ambiguous material *the contract* can initially serve either as a noun phrase (NP) complement of *understood* or as the subject of a sentential (S) complement. An unambiguous version of this sentence can be created by adding the overt complementizer *that*, as in (4b). Empirically, the underlined critical region *would be changed* is read faster in (4b) than in (4a).

The second ambiguity we investigate is the NP/Z ambiguity discussed in the introduction, and repeated here as

(5a):

(5) a. Even though the girl phoned the instructor was very upset with her for missing a lesson.
    b. Even though the girl phoned, the instructor was very upset with her for missing a lesson.

Sentences such as (5a) are referred to as NP/Z sentences because the ambiguous material *phoned* can be parsed either as a transitive verb, with the noun phrase (NP) complement *the instructor*, or as an intransitive verb with a “zero” (Z) complement. An unambiguous version of this sentence can be created by inserting a comma after the initial verb (5b); *was very upset* is read faster in (5b) than in the ambiguous (5a). This ambiguity is often perceived to be harder to resolve than NP/S.

The final type of ambiguity we study is the MV/RR ambiguity (Bever, 1970; MacDonald et al., 1992), illustrated in (6a):

(6) a. The experienced soldiers warned about the dangers conducted the midnight raid.
    b. The experienced soldiers who were warned about the dangers conducted the midnight raid.

This ambiguity is referred to as the MV/RR ambiguity because the verb *warned* can be initially parsed either as the main verb (MV) of the sentence (where the soldiers were the ones warning about the dangers) or as the verb of a reduced relative (RR) clause (where the soldiers were warned about the dangers by someone else). The MV reading is much more frequent (Fine et al., 2013), and is typically the one that is initially preferred.

The disambiguating region in the temporarily ambiguous version of each pair of sentences, underlined in the examples above, is read more slowly on average than the same region in the unambiguous version. While the first word of the disambiguating region generally disambiguates the sentence, slowdown can be observed throughout the region because of spillover (see Section 2.4). In the matched unambiguous version of each construction, these words are, of course, not disambiguating; to refer to these words in both contexts we will also use the term “critical region”.

2.2 | Self-paced reading time measurements

We focus our modeling efforts on reading times measured using the moving-window self-paced paradigm (Just et al., 1982). In this experimental paradigm, the words of each sentence are initially replaced with dashes; participants press a key to reveal the next word, at which point the previous word is replaced with dashes again. Processing difficulty on a word causes participants to take longer to advance to the next word; this slowdown often carries over to subsequent words ("spillover"; see below).

We use the publicly available self-paced reading times released by Prasad and Linzen (2019a) and Prasad and Linzen (2019b). Prasad and Linzen (2019b) had a large number of online participants recruited on Amazon Mechanical Turk (224 after standard subject exclusions) read sentences with NP/S and NP/Z ambiguities, taken from the materials of Grodner et al. (2003), where the ambiguous NP was always a plausible object of the verb (cf. Garnsey et al. 1997). They found that the average garden path effect in NP/S sentences was 15 ms, and the corresponding effect for NP/Z sentences was 28 ms. Prasad and Linzen (2019a) collected self-paced reading times for MV/RR constructions from 73 subjects on the Prolific Academic crowdsourcing platform; the mean garden path effect for this construction was 22 ms. In both studies, participants also read filler sentences, with a variety of unambiguous syntactic structures, which we use below to estimate the conversion factor between surprisal and reading time.

The effect sizes reported by Prasad and Linzen are smaller than those reported in earlier work; for comparison, Grodner et al. (2003) reported a garden path effect of 70 ms for the NP/Z ambiguity while Prasad and Linzen report an effect of around 30 ms. These differences may reflect differences between Prasad and Linzen's online participants and the in-lab participants from previous work; online experiments have obtained qualitatively similar results to earlier in-lab studies, though occasionally with faster reaction times overall (among many others, Crump et al. 2013; Enochson and Culbertson 2015; Fine and Jaeger 2016; Linzen and Jaeger 2016). The lower effect size of the replication study could also be an instance of the general finding that effect sizes reported in small-sample published studies may be exaggerated when publication is contingent on a statistically significant result, as is often the case (Vasishth et al., 2018). If Prasad and Linzen's estimates are unusually low compared to the true effect size, our results may overestimate surprisal's ability to account for the full magnitude of the garden path effect; we will return to this point below.

2.3 | Language models

A language model defines a probability distribution over sequences of words. To test the predictions of surprisal theory, we extract recurrent neural network language model surprisal—negative log probability conditioned on the preceding words—for each word in Prasad and Linzen's materials.[1] We adopt the architecture of the neural language model used by Gulordava et al. (2018). This architecture consists of two layers of long short-term memory (LSTM) recurrent units (Hochreiter and Schmidhuber, 1997). For further information about RNN language models, we refer the reader to Goldberg (2017).

Our main analyses are based on the model released by Gulordava et al. (2018); this model was trained on an 80-million word subset of English Wikipedia. We refer to this model as Wiki RNN. This particular trained model has been extensively studied in the literature, and has been shown to be sensitive to subject-verb agreement across intervening nouns (Gulordava et al., 2018), filler-gap dependencies (Wilcox et al., 2018), and constructions with temporary syntactic ambiguities (van Schijndel and Linzen, 2018; Futrell et al., 2019), among other syntactic structures.

Since Wikipedia sentences may be longer and more complex than those experienced by the average participant in the reading studies we model, an RNN language model trained on Wikipedia may assign an unrealistically large amount of probability mass to complex constructions such as the ones we investigate. This could lead to an outcome where surprisal derived from such model predicts smaller garden path effects than a model trained on simpler language data. To address this concern, we trained another RNN language model on a soap opera dialog corpus (Davies, 2011), using similar training parameters to the Wikipedia RNN above;[2] we refer to this model as Soap RNN. The average sentence length in the soap opera training corpus is 9 words, much shorter than the average sentence length of the Wikipedia training corpus (27 words), and the sentences tend to be have much simpler constructions than are typical in Wikipedia.

2.4 | Spillover

In self-paced reading, the surprisal of a word affects reading time not only at the word itself but also in at least the three subsequent words (Smith and Levy, 2013). This phenomenon, referred to as spillover, (Mitchell, 1984), has two implications: first, the garden path effect observed in human experiments is spread over multiple words; and second, reading times on the critical region are affected by the surprisal of material preceding the critical region. For concreteness, consider the MV/RR sentence (7a):

[1] Code which estimates surprisal and other incremental complexity measures from our RNN language models is available at: `https://github.com/vansky/neural-complexity.git`

[2] Our training parameters were identical except that, due to memory constraints, we used a batch size of 64 rather than the batch size of 128 used in Wiki RNN.

(7) a. The experienced soldiers warned about the dangers conducted the midnight raid.

b. The experienced soldiers who were warned about the dangers conducted the midnight raid.

Reading times on a word within the underlined critical region, such as *midnight*, are affected by spillover from other words in the critical region (e.g., *conducted*) as well as from words that precede the region (e.g., *dangers*). Ignoring the spillover from the surprisal of the words preceding the word we are currently analyzing, then, can distort our estimates of the garden path effect. Likewise, the surprisal of *midnight* affects reading times not only on *midnight* itself but also on *raid*. Consequently, restricting the analysis of reading times to the critical region, without including subsequent words, may underestimate the size of the garden path effect.

Neural network language models do not display spillover effects "out of the box": disambiguation occurs entirely at the first word of the disambiguating region (*conducted* in the above example); the subsequent words of the disambiguating region do not provide additional information about the relevant parsing decision. However, since reading times at the critical region depend on the surprisal of both critical and pre-critical words—which, in turn, is affected by the presence or absence of the phrase *who were*—linking the language model's prediction to human reading times crucially requires taking into account not only the difference across the two conditions in the surprisal of the disambiguating word itself, but also in the complex pattern of spillover influence due to surprisal. We describe our spillover-aware linking function in the following section.

2.5 | Estimating the quantitative effect of surprisal on reading times

The relationship between the surprisal of one word and reading times on a subsequent word is a non-linear function of the distance between them (Smith and Levy, 2013). We repeated the procedure described by Smith and Levy to compute these coefficients, using the reading times for **filler sentences** from Prasad and Linzen (2019b). To foreshadow the conclusions of the detailed analysis we present in this section, the total surprisal-to-RT conversion coefficient, when summed across the word itself and the three subsequent words, was approximately 2 ms/bit for all our models.[3]

To estimate the conversion coefficients, we fit a linear mixed effects model, with reading times as the dependent variable, and, as fixed effects, the following properties of the current word (w_0) and the preceding three words (w_{-3}, w_{-2} and w_{-1}): surprisal (S_{-3}, S_{-2}, S_{-1}, S_0), entropy (H_{-3}, H_{-1} , H_{-2}, H_0), entropy reduction (ΔH_{-3}, ΔH_{-2}, ΔH_{-1}, ΔH_0), word frequency (f_{-3}, f_{-2}, f_{-1}, f_0), word length (l_{-3}, l_{-2}, l_{-1}, l_0), and the position of the word in the sentence (p). Entropy and entropy reduction were computed based on the output layer of the network at each time step. In other words, we used next-word entropy rather than the entropy over all possible sequences, which is difficult to compute (for discussion, see Linzen and Jaeger 2016). We also included fixed effects for the interaction between word length and frequency within each word in the three-word spillover window (e.g., we included $f_{-1} : l_{-1}$ but not $f_{-1} : l_{-3}$). Finally, we included by-participant random intercepts. Formally, our model was as follows:

$$\begin{aligned}\mathrm{rt} \sim & S_0 + S_{-1} + S_{-2} + S_{-3} + H_0 + H_{-1} + H_{-2} + H_{-3} + \Delta H_0 + \Delta H_{-1} + \Delta H_{-2} + \Delta H_{-3} + \\ & p + l_0 * f_0 + l_{-1} * f_{-1} + l_{-2} * f_{-2} + l_{-3} * f_{-3} + (1 \mid \text{subject})\end{aligned} \tag{1}$$

where the notation $x * y$ indicates that x, y and their interaction $x : y$ were all included in the model. For our conversion, we rely on the theoretical assumption, confirmed by Smith and Levy (2013), that surprisal is linearly related to reading times. The entropy reduction hypothesis similarly predicts a linear relationship between entropy reduction and reading times; there is some empirical evidence of its efficacy in predicting reading times (Frank 2010; Linzen and Jaeger 2016; Lowder et al. 2018; for a less positive conclusion, see Aurnhammer and Frank 2019). All other predictors are the control variables used by Smith and Levy (2013).

Having regressed the reading times of filler (non-garden-path) sentences on the values of each complexity metric, we use the regression coefficient values to generate a conversion rate from each of the complexity metrics to milliseconds of reading time. Specifically, we used the learned linear combination of significant ($p < 0.01$) coefficients to predict the magnitude of garden path effects for each construction. In the case of surprisal, for example, if the S_{-1}, S_{-2}, and S_{-3} coefficients were significant, we assume that these surprisal values robustly mapped onto non-garden path reading times, and that they therefore map robustly onto garden path reading times if such reading times are driven

[3] Smith and Levy (2013) reported a total conversion coefficient of 4 ms/bit. Although their coefficient is twice the size of ours, our results still hold when we use their conversion coefficient because surprisal underestimates the empirical effect size so severely.

by surprisal effects alone. Of note, our regression analysis did not reveal a significant effect for the surprisal of the current word on its own reading time (δ_0). This finding is consistent with previous findings that spillover effects are very pronounced in self-paced reading times (e.g., Smith and Levy, 2013), and underscore the need to properly account for spillover when analyzing self-paced reading data. Overall, we predicted the spillover-corrected surprisal effect $\hat{S}(w_i)$ on the i-th word of the sentence as follows:

$$\hat{S}(w_i) = \delta_{-3}S(w_{i-3}) + \delta_{-2}S(w_{i-2}) + \delta_{-1}S(w_{i-1}) \tag{2}$$

For Wiki RNN, our estimates of the individual spillover conversion rates for surprisal were δ_{-1} = 1.1 ms/bit, δ_{-2} = 0.37 ms/bit, and δ_{-3} = 0.39 ms/bit (the full set of conversion rates, for the two language models and three complexity metrics, is given in Table 1). These conversion rates indicate, for instance, that each additional bit of surprisal of the word that occurred three words before the current word is expected to cause a slowdown of 0.39 ms on the current word. This slowdown is summed with the influence of the surprisal of the two other intervening words to produce a predicted reading time for the current word.

3 | ANALYSES

3.1 | Overview

Before we discuss our analyses, we briefly summarize the logic behind them. Recall that surprisal theory assumes that one bit of surprisal causes a fixed slowdown (an increment in milliseconds), regardless of the syntactic context in which the surprising event occurs. As such, we can measure the linear correlation between surprisal and reading times on sentences without prominent syntactic ambiguities, and use this correlation to estimate the slowdown in milliseconds caused by each bit of surprisal. If, as argued by the surprisal hypothesis, syntactic disambiguation difficulty is driven entirely by the conditional probability of the disambiguating words, this surprisal-to-RT conversion should be able to entirely explain the magnitude of the garden path disambiguation effect on the critical items in a self-paced reading study.

We report a number of analyses that rely on this logic. Analysis 1 follows the traditional analysis approach in

Measure	Model	δ_{-3}	δ_{-2}	δ_{-1}	δ_0
Surprisal	Wiki RNN	0.39	0.37	1.10	
	Soap RNN	0.44	0.91	0.83	
Entropy	Wiki RNN			-3.98	9.17
	Soap RNN			-5.50	7.04
Entropy Reduction	Wiki RNN		1.63	2.17	9.98
	Soap RNN		1.64	3.52	11.71

TABLE 1 Conversion rates for each information-theoretic measure for each RNN. We only report conversion coefficients which were determined to be significant (without correction) to $p < 0.01$ during regression to filler items. These coefficients were thought to be reliable enough to use when predicting garden path effect magnitudes in our analyses. Post hoc analysis confirmed that our results hold without this significance threshold as well.

the human behavioral literature, aggregating human reading times and model predictions over the three words of the critical region. Analysis 2 breaks down the predicted and empirical reading times for each of the words of the critical region, with the goal of determining whether language model surprisal, in conjunction with our model of spillover processing, correctly identifies the precise locus of processing difficulty in each type of ambiguity. Both Analysis 1 and Analysis 2 identify significant discrepancies between models and humans. Analysis 3 then extends the methodology of Analysis 2 to next-word entropy as well as entropy reduction computed from next-word entropy. Finally, Analysis 4 shows that the RNN language models' predictions accurately reflect the syntactic structure of temporarily ambiguous sentences, indicating that surprisal's failure to predict empirical reading times cannot be straightforwardly attributed to the language models' syntactic processing limitations.

3.2 | Analysis 1: Surprisal underestimates the magnitude of garden path effects

Using the approach described in Section 2.5, we derived spillover-adjusted reading time predictions from the Wiki RNN and Soap RNN language models. We then conducted t-tests paired by item for each combination of model

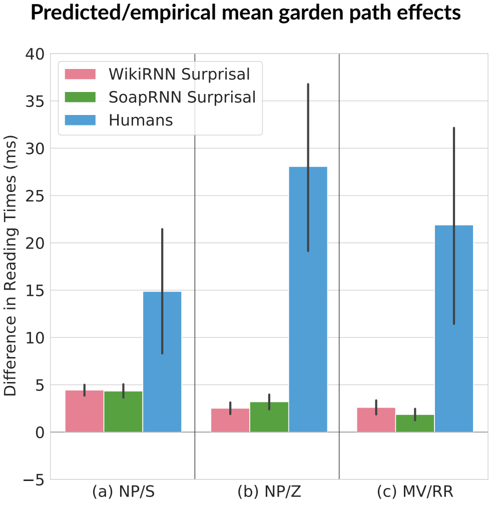

FIGURE 1 Difference in reading times between ambiguous and umambiguous sentences, averaged over the three words of the critical region, as predicted by the Wiki RNN language model (in pink) and the Soap RNN language model (in green), compared to empirical reading times on the region (in blue). Each subplot shows the disambiguation region of: (a) ambiguous NP/S sentences compared to matched unambiguous controls (example (4) in the text); (b) ambiguous NP/Z sentences compared to matched unambiguous controls (example (5) in the text); (c) ambiguous MV/RR sentences compared to matched unambiguous controls (example (6) in the text). The bars indicate the mean predicted or empirical RT, across items; the error bars represent 95% confidence intervals.

(Wiki RNN and Soap RNN) and construction (NP/S, NP/Z and MV/RR) to determine whether there was a statistically significant difference between the garden path effect predicted by the model and the empirical human effect reported by Prasad and Linzen. As shown in Fig. 1, the two models predicted effects of very similar magnitudes, and both greatly underestimated the magnitude of garden path effects across constructions; the comparison between predicted and empirical RTs was highly significant in all cases (Wiki RNN, NP/S: $p = 0.005$; Wiki RNN, NP/Z: $p < 0.001$; Wiki RNN, MV/RR: $p < 0.001$; Soap RNN, NP/S: $p = 0.006$; Soap RNN, NP/Z: $p < 0.001$; Soap RNN, MV/RR: $p < 0.001$).[4]

The conclusions of this analysis are straightforward. If the relationship between surprisal and reading times is linear, as claimed by surprisal theory—and surprisal accounts for the entire processing difficulty experienced in the disambiguation of garden path sentences—then a conversion rate derived from filler sentences, which do not exhibit perceptible syntactic ambiguities, should be able to predict reading times in garden path sentences as well; our results

[4] We report uncorrected p-values throughout the paper, and give the full list of statistical tests in the Appendix to enable post-hoc corrections for multiple comparison. In general, however, our effects are robust enough that the qualitative patterns tend to also be significant.

suggest that that is not the case.

3.3 | Analysis 2: Predicting word-by-word reading times

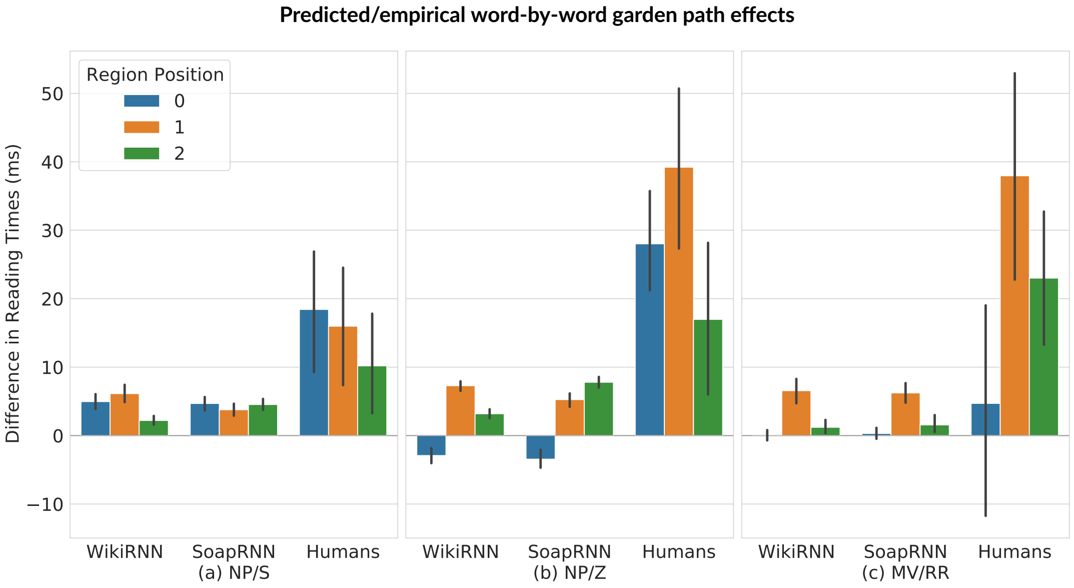

FIGURE 2 Differences in word-by-word reading times between ambiguous and unambiguous sentences on the first, second and third word of the disambiguating region, as predicted by the language models, compared to empirical reading times. The subplot shows the disambiguation region of: (a) ambiguous NP/S sentences compared to matched unambiguous controls (example (4) in the text); (b) ambiguous NP/Z sentences compared to matched unambiguous controls (example (5) in the text); (c) ambiguous MV/RR sentences compared to matched unambiguous controls (example (6) in the text). Error bars represent 95% confidence intervals.

Analysis 1 examined the garden path effect averaged over the critical region, following standard practice in the analysis of human processing of garden path sentences. To obtain a more fine-grained picture of the models' predictions, we next examine the predicted reading time *for each word* in the critical region compared with the human garden path effect observed on that word (see Fig. 2).[5]

Here too, we found that the models systematically underpredicted the empirical garden path effects in every construction and for nearly every word position. The lone exception was the first word of the disambiguating region

[5] Significance was determined using t-tests across sentence positions paired within each sentence.

of the MV/RR construction, where neither humans nor RNNs showed a significant garden path effect. Recall that if spillover is not taken into account, the first word of the disambiguating region is expected to carry the entire disambiguation effect predicted by the RNNs. We take this convergence between spillover-corrected predicted RTs and empirical RTs to be validation of our approach.

For NP/S sentences, the empirical effect is spread over the entire critical region, with no significant differences between any two points in the region (all $p > 0.05$), though there is a numerical decrease over the course of the critical region. Qualitatively, the predictions derived from Wiki RNN matched the empirical pattern, though unlike in the human data the predicted decrease in reading times did reach significance ($p < 0.001$). Soap RNN predicted a qualitatively constant effect over the entire region; significance tests showed that the effect was in fact larger on the final word in the region ($p = 0.005$), the opposite pattern from the human one, albeit with a very small effect size. Overall, the predicted time course of the NP/S effect was roughly in line with the empirical NP/S effect, though, as mentioned above, the predicted effect magnitudes are much smaller than the empirical ones.

For NP/Z sentences, the first and second words of the critical region carried the bulk of the empirical garden path response: the third word of the critical region shows a significantly smaller effect than the other two words (both comparisons $p < 0.01$). This reduction at the final word of the region was correctly predicted by Wiki RNN ($p < 0.001$). At the same time, both models predicted significant differences between all words in the region (all $p < 0.001$), and Soap RNN predicted that the effect should be *highest* in the final word of the critical region. Further, both models predicted that the first word would be read more slowly in the unambiguous condition than in the ambiguous condition (a reverse garden path effect). By contrast, humans exhibit a large NP/Z garden path effect in both the first and second word of the region.

Finally, for MV/RR sentences, the second and third words of the critical region carry the bulk of the human garden path response, with almost no empirical garden path effect observed on the first word of the critical region (the second word's effect is significantly larger than the first word's; $p < 0.0001$). Both RNNs correctly predicted that the garden path effect should be significantly larger on the second word of the region than on the first one (both $p < 0.001$). Both models also predicted that the effect should significantly subside by the third word of the region (both $p < 0.001$); the empirical effect is numerically reduced at the third word but the effect doesn't reach statistical significance. Further, the models correctly predicted that the first word in the region should not exhibit an appreciable garden path effect. As with NP/S constructions, the models were able to correctly predict the qualitative time course of the MV/RR effect throughout the region, but the magnitude of the predicted effects was much smaller than that of the empirical one.

Discussion

In human reading times, the detailed word-by-word contour of the garden path effect shows clear differences across the three constructions. This is consistent with the proposal that the disambiguation of different temporary syntactic ambiguities invokes different recovery mechanisms (compare with the distinction between "easy" and "hard" sentences of Pritchett 1988). Qualitatively speaking, the empirical time course contours of NP/S and MV/RR garden path effects were correctly predicted by both models, suggesting that humans' processing of these types of garden path sentences may be tied, through word predictability, to the frequency distributions of the syntactic constructions in questions, which are reflected in the statistics of the corpora that the RNNs were trained on. At the same time, the models predicted similar effect magnitudes for NP/S and MV/RR constructions; this contrasts with the observation that humans show a much larger effect in MV/RR constructions than in NP/S constructions. This discrepancy suggests that humans process these two constructions in different ways.

It is possible, of course, that some of the discrepancy between the empirical and predicted garden path effects arises from an incorrect estimate of the conversion rate between bits of surprisal and milliseconds of reading times. Crucially, however, the different magnitude of this discrepancy across constructions entails that even with a conversion factor large enough to predict the NP/S effect, RNNs would still underpredict the MV/RR effect (see van Schijndel and Linzen, 2018). As we discuss in the General Discussion, this result is arguably consistent with the hypothesis that the human processing of MV/RR ambiguities involves a syntactic reprocessing mechanism (Grodner et al., 2003). Such a reprocessing mechanism could amplify the effect of predictability, making it super-linear. On the other hand, the finding that RNNs were unable to predict even the qualitative time course of NP/Z garden path effects in humans supports the hypothesis that predictability-independent restructuring mechanisms are involved in recovering from this ambiguity, as proposed, among others, by Sturt et al. (1999).

In summary, as in Analysis 1, the differences between the predicted and empirical effects, in both magnitude and time course, suggest that, at a minimum, the relationship between surprisal and reading times in garden path sentences is not linear, and, more likely, that surprisal cannot on its own account for the magnitude and time course of all garden path effects in human reading.

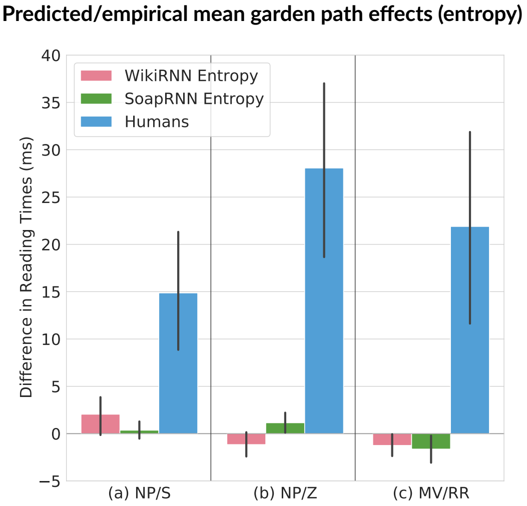

FIGURE 3 Difference in reading time between ambiguous and umambiguous sentences, averaged over the three words of the critical region, as predicted by the entropy of the Wiki RNN language model (in pink) and the Soap RNN language model (in green), compared to empirical reading times on the region (in blue). Each subplot shows the disambiguation region of: (a) ambiguous NP/S sentences compared to matched unambiguous controls (example (4) in the text); (b) ambiguous NP/Z sentences compared to matched unambiguous controls (example (5) in the text); (c) ambiguous MV/RR sentences compared to matched unambiguous controls (example (6) in the text). The bars indicate the mean predicted or empirical RT, across items; the error bars represent 95% confidence intervals.

3.4 | Analysis 3: Entropy-based complexity metrics

While much previous work has attributed garden path effects to surprisal (Hale, 2001; Levy, 2013; van Schijndel and Linzen, 2018; Futrell et al., 2019), it is not the only one-stage theory of processing difficulty proposed in the literature. In particular, two prominent information-theoretic measures have been shown to predict reading times in some contexts: single-step entropy (Roark et al., 2009; van Schijndel and Linzen, 2019) and entropy reduction (Hale, 2006; Frank, 2013; Linzen and Jaeger, 2016). To determine whether these metrics, as single-stage theories, can explain human garden path effects, we follow the same procedure we used for surprisal: we first use the filler sentences from Prasad and Linzen (2019b) to compute spillover-controlled conversion rates for each combination of model and processing difficulty metric, then use this conversion factor to predict processing difficulty in garden path sentences read by the same participants, assuming a linear relationship between the complexity metric and the observed slowdown (see Table 1).

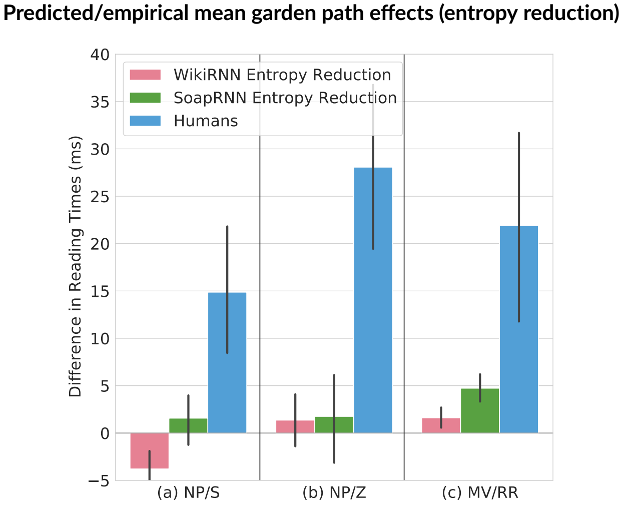

FIGURE 4 Difference in reading time between ambiguous and umambiguous sentences, averaged over the three words of the critical region, as predicted by the entropy reduction of the Wiki RNN language model (in pink) and the Soap RNN language model (in green), compared to empirical reading times on the region (in blue). Each subplot shows the disambiguation region of: (a) ambiguous NP/S sentences compared to matched unambiguous controls (example (4) in the text); (b) ambiguous NP/Z sentences compared to matched unambiguous controls (example (5) in the text); (c) ambiguous MV/RR sentences compared to matched unambiguous controls (example (6) in the text). The bars indicate the mean predicted or empirical RT, across items; the error bars represent 95% confidence intervals.

We found that entropy and entropy reduction were much poorer predictors of human garden path processing than surprisal (Figs 3 and 4). In fact, in most cases these measures predicted no effect at all (entropy reduction) or an effect in the opposite direction from the empirical one (entropy). This suggests that even if we relax the assumption that there is a linear relationship between these metrics and processing difficulty, and consider other positive and monotonic linking functions, these measures will not be able to predict human garden path effects. We stress that neither of these complexity metrics faithfully implements the Entropy Reduction Hypothesis (Hale, 2003), which requires computing entropy over complete sentences, rather than only the next word, as we did here; we are not aware of existing algorithms for estimating full-sentence entropy for RNN language models. However, our results are consistent with those of Linzen et al. (2016), who did compute full-sentence entropy from a grammar-based language model, and found that this complexity metric (unlike surprisal) did not predict a garden path effect in the correct direction.

3.5 | Analysis 4: Do RNN language models make appropriate syntactic predictions?

In the previous surprisal analyses, the language models showed qualitative garden path effects: surprisal at the critical region was higher when the region appeared in an ambiguous sentence than an unambiguous one. In most cases, however, the models' surprisal, when multiplied by the conversion coefficient we estimated, drastically underestimated the magnitude of the effect. It is possible that the models failed to predict the correct effect size simply because they were unable to take the syntactic structure of temporarily ambiguous sentences into account when making their predictions; if a language model's predictions do not match those of humans, surprisal derived from the model is unlikely to provide a good fit to human reading times. The goal of the current section is to explore the validity of this concern.

Since the RNNs made similar predictions to one another, we focus on analyzing Wiki RNN in this section. As a window into Wiki RNN's syntactic predictions at the first word of each construction's critical region, we grouped the lexical predictions of the model by the part of speech that was most frequently assigned to each of the words in the vocabulary in the Linzen et al. (2016) Wikipedia corpus. For example, although *man* can either be a noun (*see the man*) or a verb (*man the decks*), it most commonly occurs as a noun, so we would assign the probability mass associated with *man* to the noun category. Summing these probabilities over the entire vocabulary, we then inferred the model's syntactic predictions from the probability distribution it encodes over upcoming parts-of-speech, and compared it to the analogous distribution for the matched unambiguous sentence.

Results

At the beginning of the critical region of unambiguous (control) sentences, Wiki RNN assigned a high probability to the event that the upcoming word is a verb, consistent with the correct parse. This was the case for all three constructions. Conversely, in the ambiguous conditions, the model was, like humans, garden-pathed into making syntactic predictions that are not consistent with the ultimately correct parse. In particular, in ambiguous NP/S sentences (Fig. 5a), the model generally encoded the expectation that a prepositional phrase should appear next (*Mary saw the doctor at* ...); it also assigned some probability mass to the possibility that the upcoming token marks the end of the clause (i.e. a punctuation mark or a conjunction). Both of these types of continuations are consistent with the "NP" parse, where Mary saw the doctor, and not with the ultimately correct "S" parse, where Mary saw that the doctor was doing something. In ambiguous NP/Z sentences (Fig. 5b), the model again predicted that the beginning of the sentence (*When Mary visited the doctor*) would be followed by a prepositional phrase or a punctuation mark other than a period

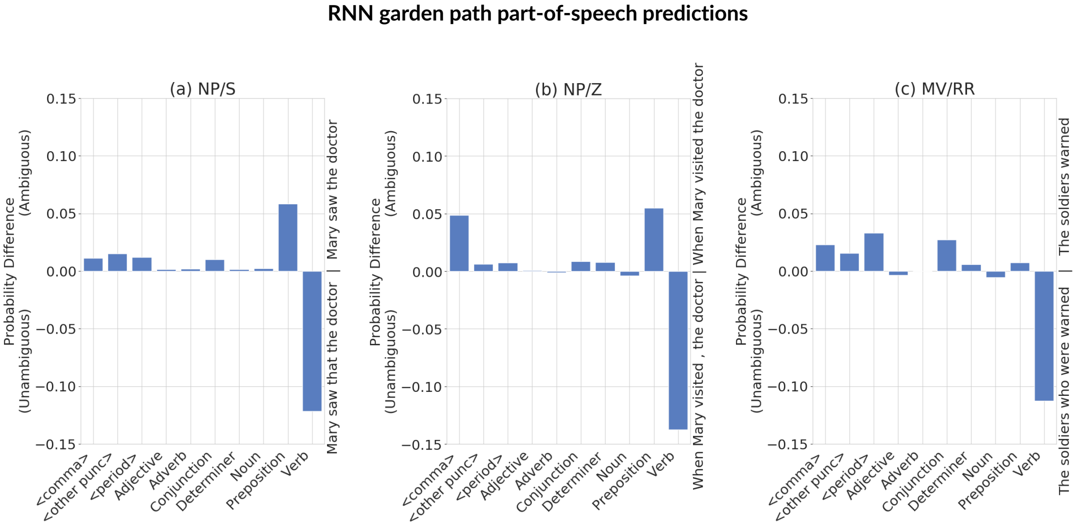

FIGURE 5 Part-of-speech predictions of the recurrent neural network language model on the first word of the critical region of unambiguous sentences, subtracted from the predictions on the same word in their ambiguous counterparts. (a) NP/S sentences; (b) NP/Z sentences; (c) MV/RR sentences.

(e.g., a comma), all continuations that are consistent with the ultimately incorrect "NP" parse, where *doctor* is the object of *visited*. Lastly, in MV/RR ambiguous sentences (Fig. 5c), the model predicted a punctuation mark would come next, most likely a period (*The soldiers warned about the dangers* .).

Overall, Wiki RNN's predictions reflect sensitivity to the syntactic structure of temporarily ambiguous sentences; this is consistent with the findings of Futrell et al. (2019). We conclude that the failure of RNN surprisal to predict the magnitude of human garden path effects cannot be attributed to the RNNs' failure to track the relevant syntactic ambiguity.

4 | GENERAL DISCUSSION

Garden path sentences are temporarily ambiguous sentences that are eventually disambiguated in favor of a dispreferred parse. In those sentences, reading times at the disambiguation point are elevated compared to matched unambiguous control sentences, commonly referred to as a garden path effect. A number of accounts have attributed this processing difficulty to reanalysis or pruning strategies specific to the human parsing system (Pritchett, 1988;

Jurafsky, 1996; Narayanan and Jurafsky, 1998; Sturt et al., 1999; Bader, 1998). More recently, proponents of the surprisal hypothesis have suggested that the elevated reading times in the disambiguating region of garden path sentences can be attributed entirely to the fact that the words in the disambiguating region are unpredictable (Hale, 2001; Levy, 2013). Since predictability affects sentence processing far beyond temporarily ambiguous sentences (Ehrlich and Rayner, 1981), such an account is preferable on parsimony grounds, as it obviates the need for assumptions that are specific to syntactic processing.

Such a parsimonious single-factor account holds an undeniable appeal. But, as we have argued, to show that word surprisal makes it unnecessary to invoke parsing-specific mechanisms in an account of garden path processing difficulty, it is not enough to show that the disambiguating word is unpredictable; rather, predictability would need to explain the full *magnitude* of the effect. Our goal in this article was to test empirically whether that is the case. To do so, we first estimated the effect of predictability on reading times in filler sentences that did not include garden path constructions. In estimating this effect, we took into account the possibility of spillover effects, where the predictability of a word affects reading times on later words.

We then estimated the surprisal of the disambiguating region in three types of garden path sentences—NP/S, NP/Z and MV/RR—from recurrent neural network (RNN) language models. We trained those models on either Wikipedia articles or soap opera dialogues. Finally, we used the global effect of surprisal on reading times to generate predicted reading times, which we then compared to empirical reading times for garden path sentences.

While the language models indeed predicted higher surprisal in the disambiguating region of temporarily ambiguous sentences compared to control sentences (in line with Hale 2001; Levy 2013; Futrell et al. 2019), the difference in surprisal systematically underpredicted the magnitude of the effect in human studies. In particular, unlike humans, which exhibit much larger garden path effects in NP/Z than NP/S sentences, language models displayed slightly *lower* surprisal in NP/Z sentences than NP/S sentences. Similarly, the language models predicted similar effect magnitudes in NP/S and MV/RR constructions, whereas in human studies MV/RR constructions show a substantially larger garden path effect than NP/S constructions. Given this complex pattern of discrepancies, then, linear functions of surprisal have no hope of deriving the human pattern, even if the true conversion coefficient between surprisal and RTs is very different from the one we estimated.

RNN surprisal is driven by simultaneous lexical predictions rather than explicitly structural predictions, unlike a syntactic parser which outputs complete descriptions of possible sentence interpretations. As a sanity check, in Analysis 4 we verified that the parts-of-speech of the words predicted by the Wiki RNN in the critical region were consistent with our expections. In the unambiguous conditions the RNN makes the correct prediction of a verb, and in the ambiguous conditions the RNN makes reasonable alternative predictions (e.g., given the context *When Mary visited the doctor*, a period indicating the end of the sentence is not assigned a significant probability, while a comma is, appropriately). These qualitative analyses indicate that the probability model within the RNN does track the expected set of syntactic parses.

4.1 | Word-by-word reading patterns in the critical region

Going beyond traditional analyses, which target mean reading times in the critical region, we explored word-by-word patterns throughout the critical region of each of the garden path constructions. In humans, this analysis revealed that the NP/S and NP/Z garden path effects are spread across the three words of the critical region, while the MV/RR garden path effect is only detectable on the second and third words of the critical region. RNN surprisal was able to predict the contour of the human garden path effect for the NP/S and MV/RR constructions, but not for NP/Z. The unique contour of the human garden path response to each construction suggests that there may be multiple distinct mechanisms that underlie each of these behavioral responses.

Two-stage accounts of human processing of garden path sentences have often hypothesized that syntactic reanalysis mechanisms rely on tree edit operations, which transform the initially preferred parse into a new parse that is compatible with the disambiguating words (Pritchett, 1988; Sturt, 1997). Under these theories, reanalysis is more costly the more the structures before and after the edit operation differ from each other. For example, Sturt et al. (1999) hypothesized that the garden path effect is larger in NP/Z than NP/S constructions because in NP/Z the ambiguous NP needs to be moved into a new subtree, which is not dominated by the subtree that contained the NP before the transformation, while in NP/S constructions the ambiguous NP remains within the same subtree throughout the reanalysis. These theories predict that the time course of processing during the critical region of garden path constructions should depend only on the similarity or dissimilarity of the associated syntactic structures, and not on the conditional probabilities of the structures in question.

Contrary to the prediction of classic reanalysis theories, we find that the word-by-word human garden path effects in NP/S and MV/RR constructions follow a similar time course to the RNN's predictions for those constructions. Since RNN predictions are solely based on the occurrence frequencies in the training data rather than reflecting human processing limitations such as working memory constraints, their ability to predict the time course of garden path processing in these constructions suggests that human reanalysis processes in these constructions is underlyingly driven by occurrence frequencies.

One repair mechanism that could produce effects such those we observed with MV/RR sentences—ones that are qualitatively consistent with the predictions of surprisal, but whose magnitudes are substantially larger than those predicted by surprisal—is the one proposed by Grodner et al. (2003). Grodner et al. hypothesized that readers suppress an initially-preferred parse once it proves to be incorrect, as in a garden path construction. The readers then reprocess the observed sequence using standard processing mechanisms but with the incorrect distractor parse suppressed. Under this theory, there are no special reanalysis mechanisms aside from a means of suppressing disconfirmed parses. This hypothesis claims that all predictability influences aside from the probability of the suppressed parse would impact both the initial parse and the subsequent reanalysis parse. As a result, this theory would predict exaggerated frequency effects whenever the reanalysis mechanism is invoked over the parallel reranking mechanism involved in surprisal theory.

4.2 | Relationship to other sources of processing difficulty

Reading behavior is affected by a range of factors other than surprisal, including word length (Just et al., 1982), dependency locality (Gibson, 2000), retrieval interference (Lewis and Vasishth, 2005), and others. To our knowledge, there are no proposals suggesting that *all* instances of syntactic processing difficulty can be attributed to surprisal; proponents of surprisal theory have proposed to supplement it with measures such as verification cost (Demberg et al., 2013) or memory and locality (Levy, 2013; Levy and Keller, 2013). Could one of these factors account for processing difficulty in garden path sentences, replacing the need for either prediction-based or reanalysis-based accounts? We believe that is unlikely, and are unfamiliar with any such proposals; in fact, factors such as word length or retrieval interference are in all likelihood perfectly matched across the ambiguous and unambiguous sentences of each garden path contrast. Those factors do affect the processing of filler sentences, which we used to estimate the conversion factor between surprisal and reading times, and we controlled for one of them (word length) when we estimated the conversion factor. This mitigates potential concerns about overestimating the influence of surprisal on reading times (see Shain, 2019). We stress, however, that our conclusions do not crucially depend on having a precise estimate of the conversion: in fact, in post hoc analysis (not presented here), we found that they held even when we doubled the conversion rate.

While discussions of garden path effects tend to focus on the differences in syntactic structure between the ambiguous and unambiguous sentences, the processing of garden path sentences is also affected by semantic plausibility—for example, the plausibility of the ambiguous NP as a direct object of the verb in NP/S sentences (Garnsey et al., 1997)—which could vary systematically between the two conditions (Padó et al., 2009). Previous studies have computed surprisal using probabilistic context-free grammar models trained on small corpora (approximately one million words). These models, while appropriately capturing the syntactic distinctions across conditions, may not capture semantic plausibility constraints very well; previous studies have proposed supplementing these models with explicit models of semantic fit (Padó et al., 2009). By contrast, in this work we computed surprisal using RNN language models trained on large corpora (e.g., 80 million words for Wiki RNN). Much previous work has shown that such language models are able to capture semantic generalizations through their distributed representations of word meaning (e.g., Mikolov et al., 2013; Levy and Goldberg, 2014), so we expect that surprisal computed by these models would be much more sensitive to semantic plausibility constraints than previous work. Similarly, numerous previous studies have found that RNN language models models learn syntactic generalizations (e.g., Gulordava et al., 2018; Wilcox et al., 2018), and in this paper we showed that their syntactic predictions in garden path sentences in particular were similar to those of humans (Analysis 3). Recent work has even shown that these models are able to encode some pragmatic inferences (e.g., Schuster et al., 2020). Taken together, we expect surprisal computed from our RNN language models to capture the combination of syntactic, semantic, and pragmatic generalizations required to account for garden path effects (Padó et al., 2009). Despite the convergence in the outcome of the prediction process, it is an open question if the mechanisms that give rise those predictions in RNNs are similar to those used by humans.

4.3 | Converging evidence for two-stage accounts from other dependent measures

Our analysis focused on self-paced reading times, a dependent measure that aggregates all sources of difficulty in language processing into a single number (the amount of time taken to read a given word). At the same time, our conclusion that predictability is insufficient to account for the strength of garden path effects is consistent with a dissociation observed in the event related potential (ERP) literature between the N400 component, which is sensitive to word predictability (Van Petten and Luka, 2012; Frank et al., 2015), and the P600 component, which is not straightforwardly related to word predictability, but is strongly modulated by disambiguation in favor of the dispreferred parse in garden path sentences (Osterhout et al., 1994). The conclusions of studies of syntactic processing that used the eye-tracking-while-reading paradigm are more mixed; while early studies found that garden path sentences are associated with a greater probability of regressive eye movements (Frazier and Rayner, 1982), it has proved difficult to isolate a consistent syntax-specific processing signature in this paradigm (Clifton Jr et al., 2007).

Unlike the experiments we presented here, which provide a direct test of surprisal's predictions at the qualitative, relative and quantitative levels, the dissociation between N400 and P600 bears on the predictions of surprisal theory only indirectly. Surprisal is intended as a computational-level theory of reading behavior, in the sense of Marr (1982); its prediction—that less predictable words should be read more slowly—can arise from the aggregated effect of any number of mental (or neural) processes. However, it is notable that the linear relationship between surprisal and reading times breaks down in the same conditions under which the N400/P600 dissociation occurs; this arguably provides converging support for a existence of a second-state reanalysis mechanism, which is indexed by the P600, and causes a slowdown in reading that is significantly more severe than predicted by surprisal (which, in turn, correlates best with the N400; Frank et al. 2015).

4.4 | Conclusion

We tested the hypothesis that word predictability can account for the full magnitude of the syntactic disambiguation difficulty that arises in three types of temporarily ambiguous sentences: NP/S, MV/RR and NP/Z. Our results do not support this hypothesis: surprisal estimated from RNN language models vastly underestimated the magnitude of the effect, especially in the MV/RR and NP/Z ambiguities. This suggests that accounts of the processing of garden path sentences may need to supplement predictability with syntactic repair mechanisms. A close inspection of the human reading times points to qualitative differences in the behavioral responses to the three constructions, again calling into question a uniform predictability-based account.

In the case of the NP/S ambiguity, we hypothesize that using neural language models trained on even larger corpora (Radford et al., 2019), or linking functions based on more sophisticated models of spillover (Shain and Schuler, 2018), may align the predictions of surprisal theory more closely with human reading times. The same could be true if the surprisal-to-milliseconds conversion rate we used was inaccurate, and needs to be replaced with a larger conversion rate. For NP/Z and MV/RR, on the other hand, the difference between the predicted and empirical effect sizes is so great that model predictions for NP/Z and MV/RR constructions would still greatly underpredict the empirical effect magnitudes even if we doubled the conversion rate. Further, when analyzed at the individual word level, the models' predictions for NP/Z sentences were so qualitatively different from the human responses that it seems unlikely that scaling up language model capacity or training corpus could mitigate the mismatch between model surprisal and humans. Finally, the empirical garden path effect magnitudes reported by (Prasad and Linzen, 2019a,b) are smaller than the garden path effect sizes reported elsewhere in the literature (e.g., the NP/Z effect reported by Grodner et al. 2003 was more than twice as large as that reported by Prasad and Linzen 2019b). If the true garden path effect is closer to that reported elsewhere in the literature, then surprisal underestimates the garden path effect to an even greater extent than we report here.

In summary, we find that single-stage processing models are unlikely to be able to accurately predict reading times of the various garden path constructions we analyzed in this work: NP/S, NP/Z, and MV/RR. At a minimum, our results indicate that the relationship between surprisal and reading times is not linear in extreme conditions such as MV/RR garden path constructions. It is possible that such a non-linear relationship may arise naturally if surprisal is augmented with the noisy channel or lossy context hypotheses (Levy, 2008; Bicknell and Levy, 2010; Gibson et al., 2013; Futrell et al., 2020). However, this possibility cannot explain the qualitatively incorrect reading time predictions we observed in NP/Z constructions. Therefore, we conclude that in addition to surprisal influences, sentence processing likely involves a syntactic repair mechanism (e.g., Sturt, 1997; Sturt et al., 1999) or reprocessing mechanism (e.g., Grodner et al., 2003) which is utilized in extreme conditions such as garden path constructions.

Acknowledgments

Thanks to Grusha Prasad, Becky Marvin, and Roger Levy for engaging and helpful discussion regarding many aspects of this project, and to Dan Grodner and Nathaniel Smith for sharing their experimental materials.

References

Aurnhammer, C. and Frank, S. L. (2019) Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, **134**, 107198.

Bader, M. (1998) Prosodic influences on reading syntactically ambiguous sentences. In *Reanalysis in sentence processing* (eds. J. Fodor and F. Ferreira), 1–56. Kluwer.

Bever, T. G. (1970) The cognitive basis for linguistic structure. In *Cognition and the Development of Language* (ed. J. R. Hayes), 279–362. New York: Wiley.

Bicknell, K. and Levy, R. (2010) Rational eye movements in reading combining uncertainty about previous words with contextual probability. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (eds. S. Ohlsson and R. Catrambone), 1142–1147. Austin, TX: Cognitive Science Society.

Clifton Jr, C., Staub, A. and Rayner, K. (2007) Eye movements in reading words and sentences. In *Eye Movements* (eds. R. P. G. van Gompel, M. H. Fischer, W. S. Murray and R. L. Hill), 341–371. Elsevier.

Crump, M. J., McDonnell, J. V. and Gureckis, T. M. (2013) Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, **8**, e57410.

Davies, M. (2011) Corpus of American Soap Operas: 100 million words. `https://www.english-corpora.org/soap`.

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

Demberg, V., Keller, F. and Koller, A. (2013) Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, **39**, 1025–1066.

Ehrlich, S. and Rayner, K. (1981) Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, **20**, 641–655.

Elman, J. L. (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, **7**, 195–225.

Enochson, K. and Culbertson, J. (2015) Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PloS one*, **10**, e0116946.

Fine, A. B. and Jaeger, T. F. (2016) The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **42**, 1362–1376.

Fine, A. B., Jaeger, T. F., Farmer, T. A. and Qian, T. (2013) Rapid expectation adaptation during syntactic comprehension. *PLOS ONE*, **8**, e77661.

Frank, S. (2013) Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, **5**, 475–494.

Frank, S. L. (2010) Uncertainty reduction as a measure of cognitive processing effort. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (ed. J. T. Hale), 81–89. Uppsala, Sweden: Association for Computational Linguistics.

Frank, S. L., Otten, L. J., Galli, G. and Vigliocco, G. (2015) The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, **140**, 1–11.

Frazier, L. (1979) *On comprehending sentences: Syntactic parsing strategies*. Ph.D. thesis, University of Connecticut.

Frazier, L. and Fodor, J. D. (1978) The sausage machine: A new two-stage parsing model. *Cognition*, **6**, 291–325.

Frazier, L. and Rayner, K. (1982) Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, **14**, 178–210.

Futrell, R., Gibson, E. and Levy, R. P. (2020) Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, **44**, e12814.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M. and Levy, R. (2019) Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 32–42. Minneapolis, Minnesota: Association for Computational Linguistics.

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

Gibson, E. (2000) The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, 95–126. Cambridge, MA: MIT Press.

Gibson, E., Bergen, L. and Piantadosi, S. T. (2013) Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, **110**, 8051–8056.

Gibson, E. A. F. (1991) *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Ph.D. thesis, Carnegie Mellon University.

Goldberg, Y. (2017) *Neural network methods for natural language processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Goodman, J. T. (2001) A bit of progress in language modeling. *Computer Speech & Language*, **15**, 403–434.

Gorrell, P. (1995) *Syntax and Parsing*. Cambridge University Press.

Grodner, D. J., Gibson, E., Argaman, V. and Babyonyshev, M. (2003) Against repair-based reanalysis in sentence comprehension. *Journal of Psycholinguistic Research*, **32**, 141–166.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T. and Baroni, M. (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)*, 1195–1205. Association for Computational Linguistics.

Hale, J. (2001) A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8. Pittsburgh, PA: Association for Computational Linguistics.

— (2003) The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, **32**, 101–123.

Hale, J. (2006) Uncertainty about the rest of the sentence. *Cognitive Science*, **30**, 609–642.

Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural Computation*, **9**, 1735–1780.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N. and Wu, Y. (2016) Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

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

Just, M. A., Carpenter, P. A. and Woolley, J. D. (1982) Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, **111**, 228–238.

Levy, O. and Goldberg, Y. (2014) Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308. Baltimore, Maryland: Association for Computational Linguistics.

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

Levy, R. (2008) A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243. Stroudsburg, PA: Association for Computational Linguistics.

— (2013) Memory and surprisal in human sentence comprehension. In *Sentence Processing* (ed. R. P. G. van Gompel), 78–114. Hove: Psychology Press.

Levy, R. P. and Keller, F. (2013) Expectation and locality effects in german verb-final structures. *Journal of Memory and Language*, **68**, 199–222.

Lewis, R. L. and Vasishth, S. (2005) An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, **29**, 375–419.

Linzen, T., Dupoux, E. and Goldberg, Y. (2016) Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, **4**, 521–535.

Linzen, T. and Jaeger, T. (2016) Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, **40**, 1382–1411.

Lowder, M. W., Choi, W., Ferreira, F. and Henderson, J. M. (2018) Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, **42**, 1166–1183.

MacDonald, M. C., Just, M. A. and Carpenter, P. A. (1992) Working memory constraints on the processing of ambiguity. *Cognitive Psychology*, **24**, 56–98.

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*, vol. 2. New York: Freeman.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J. and Khudanpur, S. (2010) Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, 1045–1048. Makuhari, Chiba, Japan.

Mikolov, T., Yih, W.-t. and Zweig, G. (2013) Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. Atlanta, Georgia: Association for Computational Linguistics.

Mitchell, D. C. (1984) An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In *New Methods in Reading Comprehension Research* (eds. D. E. Kieras and M. A. Just), 69–89. Hillsdale, NJ: Erlbaum.

Narayanan, S. and Jurafsky, D. (1998) Bayesian models of human sentence processing. In *Proceedings of the twelfth annual meeting of the cognitive science society*.

Osterhout, L., Holcomb, P. and Swinney, D. (1994) Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**, 786–803.

Padó, U., Crocker, M. W. and Keller, F. (2009) A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, **33**, 794–838.

Prasad, G. and Linzen, T. (2019a) Do self-paced reading studies provide evidence for rapid syntactic adaptation? *PsyArXiv preprint PsyArXiv:10.31234/osf.io/9ptg4*.

— (2019b) How much harder are hard garden-path sentence than easy ones? *OSF preprint osf:syh3j*.

Pritchett, B. L. (1988) Garden path phenomena and the grammatical basis of language processing. *Language*, **64**, 539–576.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) Language models are unsupervised multitask learners. *OpenAI Blog*, **1**, 9.

Roark, B., Bachrach, A., Cardenas, C. and Pallier, C. (2009) Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *Proceedings of EMNLP*, 324–333.

Schuster, S., Chen, Y. and Degen, J. (2020) Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5387–5403. Online: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.acl-main.479.

Shain, C. (2019) A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4086–4094. Minneapolis, Minnesota: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/N19-1413.

Shain, C. and Schuler, W. (2018) Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of EMNLP 2018*. Association for Computational Linguistics.

Smith, N. J. and Levy, R. (2013) The effect of word predictability on reading time is logarithmic. *Cognition*, **128**, 302–319.

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

Sturt, P. (1997) *Syntactic reanalysis in human language processing*. Ph.D. thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, Scotland.

Sturt, P. and Crocker, M. W. (1996) Monotonic syntactic processing: A cross-linguistics study of attachment and reanalysis. *Language and Cognitive Processes*, **11**, 449–494.

Sturt, P., Pickering, M. J. and Crocker, M. W. (1999) Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, **40**, 136–150.

Taylor, W. L. (1953) "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*.

Van Petten, C. and Luka, B. (2012) Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, **83**, 176–190.

van Schijndel, M. and Linzen, T. (2018) Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*. Cognitive Science Society.

— (2019) Can entropy explain successor surprisal effects in reading? In *Proceedings of the 2nd Annual Meeting of the Society for Computation in Linguistics (SCiL)* (eds. G. Jarosz and J. Pater). New York, NY: Society for Computation in Linguistics.

van Schijndel, M., Schuler, W. and Culicover, P. W. (2014) Frequency effects in the processing of unbounded dependencies. In *Proc. of CogSci 2014*. Cognitive Science Society.

Vasishth, S., Mertzen, D., Jäger, L. A. and Gelman, A. (2018) The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, **103**, 151–175.

Wilcox, E., Levy, R., Morita, T. and Futrell, R. (2018) What do RNN language models learn about filler-gap dependencies? In *Proceedings of BlackboxNLP*.

A | APPENDIX

A.1 | Statistical tests

A.1.1 | Analysis 1: Mean garden-path effects

We conducted 12 t-tests of whether the mean spillover-corrected predictions differed from the mean human reading times (paired by item) or from 0 (1-sample). 2 models × 3 constructions × 2 comparisons.

A.1.2 | Analysis 2: Spillover garden path contours

For each of the two spillover-corrected model predictions and the human responses, we conducted 27 t-tests of whether each word in the critical region of a construction differed from each other word (paired by item) or 0 (1-sample). 3 word positions × 3 constructions × 3 comparisons.

A.1.3 | Analysis 3: Entropy-based mean predictions

We conducted 12 t-tests of whether the mean spillover-corrected predictions differed from the mean human reading times (paired by item) or from 0 (1-sample). 2 models × 3 constructions × 2 comparisons.