REGULAR ARTICLE

Journal Section

Neural network surprisal predicts the existence but not the magnitude of human 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 Cognitive Science, Johns Hopkins University, Baltimore, MD, 21218, 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 estimate word-by-word surprisal using a recurrent neural network language model, and compare those estimates to human behavioral data. While predictability did predict the magnitude of easier garden path sentences (NP/S constructions), it underestimated the disambiguation difficulty for harder garden path sentences (NP/Z and MV/RR). These results support two-stage models in which recovery mechanisms beyond predictability are involved in processing hard garden path sentences.

KEYWORDS

self-paced reading, garden path, neural networks

1 | INTRODUCTION

The beginning of a sentence is often compatible with multiple parses, and readers tend to 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 in which the girl phoned the instructor; in other words, where *the instructor* is the direct object of *phoned*. Yet that interpretation leaves no viable subject for the subsequent verb, *was*. The realization that the initially dispreferred parse is in fact the correct one leads to elevated reading times at the disambiguating region *was very upset*, compared to the same words 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*, ensuring that readers do not 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 the critical region between (1) and (2) as a *garden path effect* (Bever, 1970).

Garden path effects have motivated cognitive theories in which readers only consider one of the possible partial parses of the sentence (Frazier and Fodor, 1978; Pritchett, 1988) or a subset of those parses (Gibson, 1991; Jurafsky, 1996) at each point in the sentence. In those theories, processing difficulty results from the reanalysis required to reconstruct the correct parse that was discarded or was not considered in the first place (Pritchett, 1988; Gorrell, 1995; Sturt and Crocker, 1996; Sturt, 1997; Bader, 1998).

An alternative proposal, which we will refer to as the surprisal hypothesis (Hale, 2001; Levy, 2008), assumes a fully parallel parser. This parser does not discard dispreferred parses; it simply associates a lower probability with those parses compared to the preferred parse. 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): the word *was*, which is only consistent with a low-probability parse, is read more slowly in the ambiguous (1) than the unambiguous (2) simply because it is less predictable in that context (Hale, 2001; Levy, 2013). If it is consistent with the empirical data, the surprisal hypothesis is preferable to two-stage models on parsimony grounds: 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 additional reanalysis mechanisms that come into play only in the disambiguation of temporarily ambiguous sentences.

1.1 | Testing the surprisal account of garden path effects

The goal of this paper is to test the hypothesis that disambiguation difficulty can be reduced to word predictability, and to do so in a more systematic way than has been done before. What is in question is not whether *was* is less predictable in the temporarily ambiguous (1) than in the unambiguous (2); that is a plausible hypothesis and one that has been demonstrated in prior work (e.g., Levy 2013). We argue that such a difference in predictability does not in itself justify the claim that garden path effects can be *reduced* to predictability. In particular, if surprisal is the only driver

of disambiguation difficulty, word predictability would need to account for the relative difference in disambiguation difficulty across different types of syntactically ambiguous sentences. The following sentence, for example, is similar to (1) in that a noun phrase initially interpreted as a direct object of a verb needs to be reanalyzed as a subject:

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

Despite this superficial similarity, disambiguation in sentences such as (3), referred to as "NP/S" sentences, does not lead to the same degree of processing difficulty as in "NP/Z" sentences such as (1) (Pritchett, 1988; Sturt et al., 1999). In the absence of the special reanalysis mechanisms that have been proposed account for this difference (Pritchett, 1988; Bader, 1998), the surprisal hypothesis can only account for the greater disambiguation difficulty if *would* in (3) can be shown to be more predictable (compared to a control) than *was* is in (1).

Surprisal would also need to account for the absolute *magnitude* of the garden path effect in each case. The surprisal hypothesis predicts that the same (logarithmic) relationship between predictability and reading time should hold in temporarily ambiguous sentences and unambiguous ones. Smith and Levy (2013) report that each unit ("bit") of surprisal leads to a slowdown of approximately 4 ms in self-paced reading experiments. The garden path effects reported in the literature are often on the order of magnitude of dozens of milliseconds (e.g., 70 ms for NP/Z sentences such as (1) in Grodner et al. 2003). For surprisal to explain this difference, then, the conditional probability of *was* in (1) would need to be 2^{18} times lower than in (2).[1]

1.2 | Estimating surprisal

How can we obtain quantitative estimates of the predictability of a word? Traditionally, predictability estimates were obtained by having participants 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 be estimated as the proportion of participants who completed the fragment with that word. While this method is effective for distinguishing highly predictable words (e.g., $P(w) = 0.8$) from moderately predictable words ($P(w) = 0.1$), it is not effective for making distinctions among lower probability words, for the following reason: even if we assume that participants in the cloze experiment occasionally consider the dispreferred parse, millions of

[1] 70 ms / 4 ms per bit $\approx$ 18, so *was* in (1) must produce 18 times more bits (2^1) of surprisal than *was* in (2), or equivalently must be 2^{18} times less likely, to produce a 70 ms effect on reading times.

participants would be required to accurately estimate probabilities on the order of magnitude of 2^{-18} (as needed to model the relevant NP/Z continuations).

Instead, recent work estimates the predictability of words using probabilistic *language models*, computational models that use a large training corpus to define probability distributions over sequences of words (Goodman, 2001). 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, syntactically complex sentences such as (1) require 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.

1.3 | Overview of experiments

To test the surprisal account of garden path effects, we modeled the results of self-paced reading studies from the existing literature, using surprisal estimates derived from an RNN language model. Participants in the experiments read NP/Z sentences such as (1), NP/S sentences such as (3), and sentences with ambiguous reduced relative clauses, modeled after the classic ambiguity *the horse raced past the barn fell* (MV/RR, Bever 1970). To derive reading time predictions from our language models, we used published correlations between reading times and surprisal, paying close attention to the possible effect of spillover (processing difficulty on an earlier word affecting reading times on a later word).

To anticipate our results, RNN surprisal correctly predicted that disambiguation should lead to a slowdown compared to unambiguous controls in all three constructions. Yet it underestimated the magnitude of the slowdown when effect of predictability was estimated using unambiguous words. The discrepancy differed across constructions: it was small in NP/S, moderate in MV/RR, and very large in NP/Z. In fact, RNN surprisal predicted numerically larger dis-

ambiguation difficulty in NP/S than NP/Z sentences, the opposite pattern from humans. 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 temporary syntactic ambiguities (Frazier, 1979). The first type is the NP/S ambiguity, illustrated in (4a):

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

b. The employees understood that the contract would be changed very soon to accommodate all parties.

This ambiguity is referred to as NP/S, because *the contract* can initially serve either as a noun phrase (NP) complement to *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 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 verb *phoned* is initially either transitive, with the noun phrase (NP) complement *the instructor*, or intransitive, 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 MV/RR because the verb *warned* can be initially perceived as either as a main verb construction (MV, in which the soldiers were warning about the dangers) or as the less frequent reduced relative construction (RR, in which the soldiers were warned about the dangers).

The disambiguating region in the temporarily ambiguous version of each pair of sentences is read more slowly on average than the same region in the unambiguous version. The critical region is typically taken to be *would be changed* in the NP/S case (4), *was very upset* in the NP/Z case (5), and *conducted the midnight* in the MV/RR case (6).

2.2 | Self-paced reading times

We focus our modeling efforts on moving-window self-paced reading times (Just et al., 1982), which are commonly used to study sentence processing. 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").

We use the publicly available self-paced reading times released by Prasad and Linzen (2019b) and Prasad and Linzen (2019a). Prasad and Linzen (2019b) had a large number of participants (224 after standard subject exclusions) recruited on Amazon Mechanical Turk read sentences with NP/S and NP/Z ambiguities. 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.

2.3 | Language models

A language model defines a probability distribution over sequences of words. To test the predictions of surprisal theory, we extract language model surprisal (negative conditional log probability) on each word in the Prasad and Linzen stimuli conditioned on the preceding words.[2] We used language models based on two computational architectures: recurrent neural networks and probabilistic context free grammars.

Recurrent neural networks

We used the RNN language model trained by Gulordava et al. (2018) on 80 million words of English Wikipedia. This particular trained model has been shown to be sensitive to syntactic dependencies, including subject-verb agreement across intervening nouns (Gulordava et al., 2018), filler-gap dependencies (Wilcox et al., 2018) and temporary syntactic ambiguity (van Schijndel and Linzen, 2018; Futrell et al., 2019). This model had two long short-term memory (LSTM) layers (Hochreiter and Schmidhuber, 1997) with 650 hidden units each, 650-dimensional word embeddings, a dropout rate of 0.2, a batch size of 128, and was trained for 40 epochs (with early stopping).

Grammar-based

Alongside RNN language models, we experiment with language models based on probabilistic context-free grammars (PCFGs; Stolcke, 1995; Roark, 2001), following previous work on modeling temporarily ambiguous sentences (Hale, 2001; Levy, 2013; Linzen and Jaeger, 2016). Grammar-based models require manually parsed corpora for training; such corpora are necessarily smaller and may not provide evidence for the fine-grained lexicosyntactic patterns that underlie some of the garden path effects (Jurafsky, 1996). At the same time, the fact that they construct explicit syntactic structure may counteract their deficits in linguistic accuracy.

In preliminary work (van Schijndel and Linzen, 2018), we used three probabilistic context-free grammar (PCFG) parsers to predict the garden path effects reported by Grodner et al. (2003): a top-down parser (Roark, 2001), a left-corner parser (van Schijndel et al., 2013), and the same left-corner parser when trained using a categorial grammar (Nguyen et al., 2012). In the present study we focus on the PCFG parser which best predicted the garden path effects in that work: the categorial grammar-trained parser of van Schijndel et al. (2013). This left-corner parser was trained on

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

categorial grammar annotations (Nguyen et al., 2012) of the Wall Street Journal corpus (Marcus et al., 1993). This annotation produces a high degree of context-sensitivity reflecting deep syntactic dependencies similar to head-driven phrase structure grammar (HPSG; Pollard and Sag, 1994). Additional context-sensitivity is obtained by applying three iterations of the Petrov et al. (2006) split-merge-smooth procedure, which splits coarse syntactic categories such as noun phrases into finer-grained ones based on distributional patterns in the corpus.

3 | ANALYSES

3.1 | Overview

We report a number of analyses relating language model surprisal to reading times. In Analysis 1, we convert surprisal values derived from language models into predicted reading times on the disambiguating region, and compare the magnitude of those predictions with garden path effects found in human experiments. Analysis 2 more closely investigates the processing of each kind of garden path construction by comparing predicted and empirical reading times for each of the words of the critical region. Finally, Analysis 3 shows that the RNN language model's predictions accurately reflect the syntactic structure of temporarily ambiguous sentences, indicating that surprisal's failure to predict empirical reading times cannot be attributed to the RNN LM's syntactic processing limitations.

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

Surprisal theory predicts that one bit of surprisal should cause the same amount of slowdown regardless of the syntactic context in which the surprising event occurs. Therefore, we can use the observed linear correlations between surprisal and reading times in a broad-coverage reading time corpus 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.

In order to obtain a surprisal-to-RT conversion factor that predicts reading times in a range of linguistic contexts, we used the reading time data and analysis of Smith and Levy (2013). They demonstrated a linear relationship between surprisal and self-paced reading times collected as 33 participants read the Brown corpus of American English (Francis

and Kucera, 1979).

Modeling spillover

Smith and Levy (2013) showed that the surprisal of a word affected reading time not only at the word itself but also in at least the three subsequent words ("spillover"; Mitchell, 1984). The relationship between surprisal and reading times differs in magnitude for each of these four words. We therefore apply the individual conversion rate (δ_i) for each spillover word (w_i) separately to obtain a reading time prediction for each individual word in the critical region; this prediction incorporates the effect of spilled over surprisal from previous words:

$$\hat{S}(w_i) = \delta_{-3}S(w_{i-3}) + \delta_{-2}S(w_{i-2}) + \delta_{-1}S(w_{i-1}) + \delta_0 S(w_i) \tag{1}$$

We estimated the individual spillover conversion rates (δ_{-3} through δ_0) from the Smith and Levy (2013) data (shown in Fig. B1b in their paper): $\delta_0 = 0.53$ ms/bit, $\delta_{-1} = 1.53$ ms/bit, $\delta_{-2} = 0.92$ ms/bit, and $\delta_{-3} = 0.84$ ms/bit. 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 0.84 ms on the current word. This slowdown is summed with the influence of the surprisal of the two other intervening words, as well as the influence of surprisal of the current word, to produce a predicted reading time for the current word.

Comparing the predicted and empirical effects

We conducted two-sample t-tests within each construction to determine whether there was a statistically significant difference between the predicted garden path effect and the empirical effect reported by Prasad and Linzen. We performed a conservative Bonferonni correction for multiple comparisons; since this paper includes 57 significance tests (see Appendix for the full list), we used a significance threshold of $p = 0.00087$ uncorrected.

The spillover-controlled model predictions are shown in Fig. 1. When model predictions were converted to reading times in this way, both models' mean predictions for the NP/S garden path effect were not significantly different from the empirical effect (RNN $p > 0.1$, PCFG $p > 0.04$). The predictions for the other constructions, however, significantly underestimated the magnitude of the garden path effect.

As a sanity check, we conducted one-sample t-tests assessing whether the mean surprisal over the critical region was significantly larger than zero for each model-construction pair (see Fig. 2). The difference in surprisal was sig-

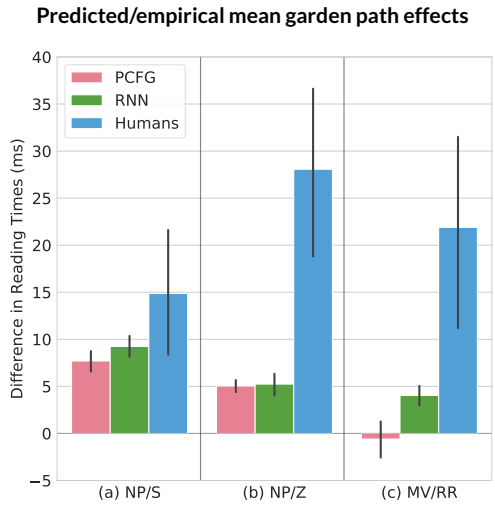

FIGURE 1 Difference in reading time between ambiguous and umambiguous sentences, averaged over the three words of the critical region, as predicted by the language models when spillover is taken into account for each word (in pink and green), compared to empirical reading times on the region (in blue). 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.

nificant for all three constructions in both language models, with the exception of the PCFG in MV/RR constructions ($p = 0.03$ uncorrected). These results indicate that language model surprisal, especially from an RNN trained on a large amount of data, is generally sufficient to predict the *existence* of garden path effects in reading (Hale, 2001). However, surprisal from neither model adequately predicted the *magnitude* of the garden path effect in reading times.

Comparison across constructions

We conducted two-sample t-tests to test whether the three constructions differed in the excess surprisal of each model on the disambiguating region (in ambiguous compared to unambiguous sentences). For the RNN, excess surprisal in NP/Z sentences was significantly higher than excess surprisal in MV/RR sentences (see Fig. 2) with no other significant differences between constructions.[3] This resembles the empirical data where the garden path effect is largest in NP/Z sentences. However, the magnitude of the empirical garden path effect in MV/RR sentences is interme-

[3] The PCFG predicted difference between NP/Z and MV/RR sentences just missed significance with $p = 0.0009$.

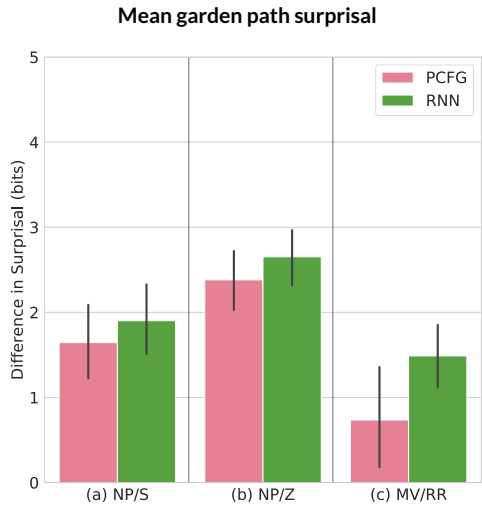

FIGURE 2 Difference between the averaged surprisal across the disambiguating region of garden path sentences compared to the averaged surprisal of the same words in matched unambiguous controls. The bar plots show the mean prediction across items; the error bars represent 95% confidence intervals. (a) NP/S sentences (example (4) in the text). (b) NP/Z sentences (example (5)). (c) MV/RR sentences (example (6) in the text).

diate between that of NP/S and NP/Z sentences, whereas our models predicted a numerically smaller effect in MV/RR sentences than in the other constructions.

When we scaled reading times using our spillover-controlled conversion rate, the scaled surprisal from both models in NP/S sentences became significantly larger than the scaled surprisal in both other constructions (see Fig. 1). Since the empirical garden path effect was smallest in NP/S sentences, our results indicate that controlling for spillover widens the gap between surprisal and empirical garden path effects in reading times, suggesting again that an additional factor beyond surprisal contributes to garden path effects in reading.

Parse failures

An analysis of the parses produced by the PCFG reveals that in many cases the parser failed to recover the intended parse by the end of the sentence.[4] Failure to recover the intended parse could indicate that the correct reading was pruned from the available hypotheses before the disambiguation point. If that happens, the model's surprisal would not accurately represent the difference in predictability between the correct and incorrect interpretations, since the

[4] Failure rate by construction: NP/S 35%, NP/Z 30%, MV/RR 50%.

correct interpretation was not tracked by the model. Nevertheless, despite the large number of parse failures, which were especially common for MV/RR constructions, removing items containing parse failures had little effect on our results. For an analysis of the results excluding parse failures, see the Supplementary Materials.

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

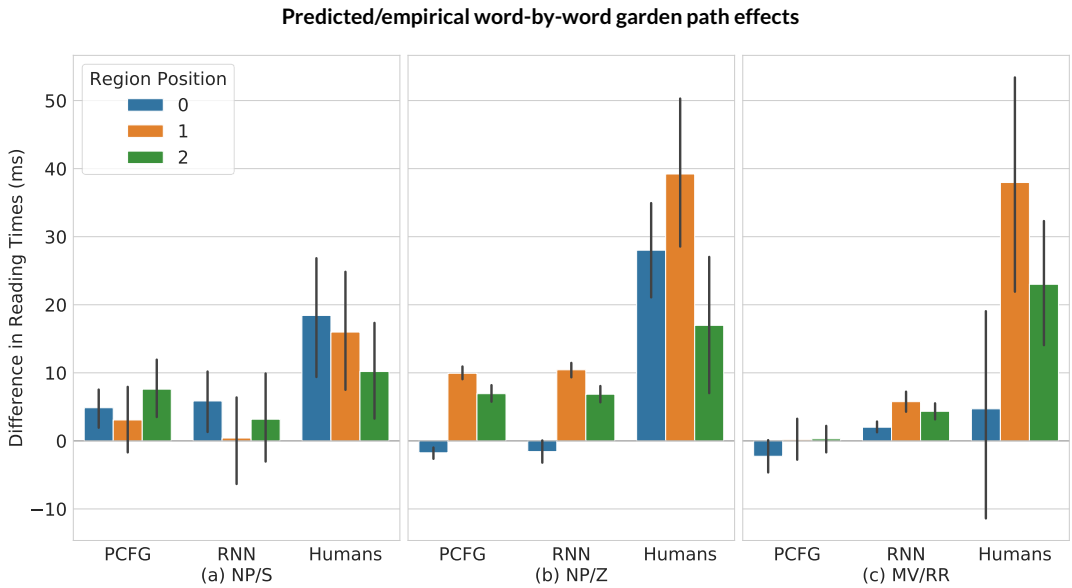

FIGURE 3 Differences in word-by-word reading times between ambiguous and unambiguous sentences as predicted by the language models when spillover is accounted for each word, compared to empirical reading times. Each subplot shows the disambiguation region of: (a) ambiguous NP/S sentences compared to 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 bar plots show the mean predictions across items at word 0, word 1, and word 2 of the disambiguating region; the 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 response for that word (see Fig. 3).

For NP/S sentences, both the predicted and the empirical effect were spread evenly over the disambiguation region: the effect on each word did not significantly differ from the effect on any other word according to pairwise t-tests (PCFG predictions, RNN predictions and empirical reading times all $p > 0.05$). At the same time, the word-by-word analysis suggests that both models numerically underestimated the disambiguation difficulty experienced by humans at every word; this contrasts with Analysis 1, in which the RNN's predictions did not significantly differ from the empirical reading times in NP/S sentences.

For NP/Z sentences, the predicted and empirical effect occurred primarily on the second word of the region and then tapered off. The models predicted no garden path effect at the initial disambiguating word, whereas humans exhibited a large garden path effect at that point. This difference is a by-product of our spillover controls; before accounting for spillover the primary difference in surprisal in the critical region was at the initial word. Accounting for spillover not only spreads the effect of surprisal over the entire critical region, but also spreads the absence of a surprisal difference from the words preceding the critical region into the first word of the region (see the Supplementary Materials for word-level plots which do not control for spillover). While the human NP/Z garden path effect appeared on every word in the region, differences in slowdown across the three words did not reach significance due to high variability.

The failure of the PCFG LM to predict the MV/RR garden path effect on the critical region as a whole was not a consequence of averaging: the present analysis shows that this failure extends to each of the words of the region. By contrast, while the humans and the RNN LM showed little effect at the initial disambiguating word, there was a large effect at the second word of the critical region, after which there was a non-significant lessening of the garden path effect on the final word. For the RNN, the qualitative MV/RR pattern was significant, but the human reading times do not differ significantly across the three words due to high variability.

Discussion

In humans, the qualitative differences in the contour of the garden path effect across the three constructions is consistent with the proposal that each construction invokes a distinct recovery mechanism (compare the Pritchett 1988 distinction of "easy" versus "hard" sentences). In particular, the NP/S effect could plausibly be an example of the surprisal effect seen more broadly; the NP/Z effect is similar in its word-by-word pattern to the NP/S effect, except its magnitude is much larger; and the MV/RR pattern is distinct from the other two constructions in that there is no garden path effect at the first word of the critical region.

The predictions we derived from the RNN, based entirely on surprisal, followed the same pattern as the distinctive human MV/RR garden path effect, albeit at a much smaller magnitude. This result may be taken to indicate that there is some syntactic repair mechanism at work in human processing of MV/RR ambiguities, which exaggerates the influence of predictability. As we explain in the discussion, Grodner et al. (2003) hypothesized just such a mechanism. On the other hand, since the RNN is unable to predict the time course of NP/Z processing in humans, we tentatively conclude that there may be some predictability-independent repair mechanism at work, such as the syntactic repair mechanisms hypothesized by Sturt et al. (1999). Overall, we conclude that surprisal is unlikely to account for the magnitude and time course of all garden path effects in human reading.

3.4 | Analysis 3: Does the RNN language model make appropriate syntactic predictions?

In the previous analyses, the language models showed qualitative garden path effects: surprisal at the critical region was higher when the region appeared in an ambiguous rather than unambiguous sentence. In most cases, however, the models' surprisal, when multiplied by the slowdown factor we estimated from Smith and Levy (2013), drastically underestimated the magnitude of the effect. One potential explanation is that the models failed to take the syntactic structure of temporary 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 address this concern.

As this concern is particularly relevant for RNN language models, which do not explicitly parse the sentence, in this section we focus on analyzing the RNN's syntactic predictions. As a window into the 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, while *man* can 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.

Results

In each unambiguous condition, the model assigned much more probability mass to the (ultimately correct) possibility that the upcoming word would be a verb than in the ambiguous conditions. Conversely, in the ambiguous conditions the model was garden-pathed into making incorrect syntactic predictions that align with those that humans are likely

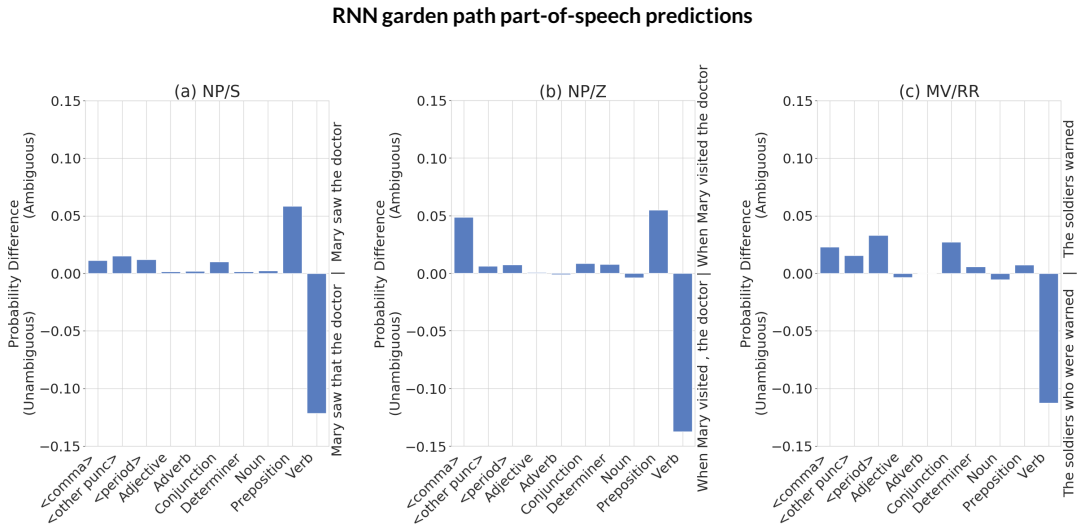

FIGURE 4 Part-of-speech predictions of the recurrent neural network language model on the first word of the critical region of unambiguous sentences, subtracted from the predictions on the same word in their ambiguous counterparts. (a) NP/S sentences; (b) NP/Z sentences; (c) MV/RR sentences.

to make. In particular, in ambiguous NP/S sentences (Fig. 4a), the model generally expected a prepositional phrase to come next (*Mary saw the doctor at* …), followed by a lower probability expectation that a clause conclusion (i.e. punctuation or a conjunction) will come next. In ambiguous NP/Z sentences (Fig. 4b), the model predicted that the beginning of the sentence (*When Mary visited the doctor*) will be followed by a prepositional phrase or a punctuation mark other than a period (e.g., a comma). Lastly, in the MV/RR ambiguous condition (Fig. 4c), the model predicted a punctuation mark will come next, most likely a period (*The soldiers warned about the dangers* .). Overall, the RNN LM's predictions reflect sensitivity to the syntactic structure of temporary ambiguous sentences (Futrell et al., 2019). We conclude that the reduced magnitude of the RNN LM's surprisal compared to human RTs cannot be attributed to the RNN's failure to track the relevant syntactic ambiguity.

4 | DISCUSSION

Garden path sentences are temporarily ambiguous sentences that are eventually disambiguated in favor of a dispreferred parse. This leads to elevated reading times at the disambiguation point compared to matched control sentences

(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 specific to parsing.

Despite the undeniable appeal of such a parsimonious single-factor account, we have argued that for word surprisal to obviate the need for parsing-specific mechanisms in accounting for processing difficulty in garden path sentences, it would not be enough to show that the disambiguating word is unpredictable; 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, including spillover effects, from a study that used a corpus composed primarily of unambiguous sentences (Smith and Levy, 2013). We then estimated the surprisal of the disambiguating region in three types of garden path sentences—NP/S, NP/Z and MV/RR—from two language models, one using a probabilistic context-free grammar (PCFG) and one using a recurrent neural network (RNN). Based on the overall effect of surprisal on reading times reported by Smith and Levy, we converted those surprisal estimates into reading time predictions, which we then compared to empirical reading times from human studies.

While the language models indeed predicted higher surprisal in the disambiguating region temporary ambiguous sentences compared to control sentences (in line with Hale 2001; Levy 2013; Futrell et al. 2018), the difference in surprisal systematically underpredicted the magnitude of the effect in human studies. In particular, while humans exhibit much larger garden path effects in NP/Z than NP/S sentences, language model surprisal was slightly *lower* in NP/Z sentences when spillover was accounted for.

Detailed reading patterns in the critical region

Going beyond modeling mean reading times over the critical region, we explored word-by-word responses throughout the critical region of each of these constructions. In humans, this analysis revealed that the NP/S and NP/Z garden path effects are distributed across the critical region, while in MV/RR sentences there is no significant garden path effect on the first word of the critical region. RNN surprisal was able to predict the contour of the empirical NP/S and MV/RR garden path effect but not the NP/Z effect. The unique contour of the human garden path response to each

construction suggests that there may be multiple distinct mechanisms that underlie each of those responses.

Some theories of garden path processing have hypothesized that syntactic reanalysis mechanisms use edit operations to directly manipulate incorrect syntactic structures (as opposed to merely increasing the probability of the correct structures; Pritchett 1988; Sturt 1997). Under these theories, reanalysis should take more time when the two structures involved in the conversion are less similar to one another. And that extra time would produce a larger garden path effect. For example, Sturt et al. (1999) hypothesized that the NP/Z garden path effect is larger than the NP/S garden path effect because the ambiguous NP/Z NP subtree must be moved into a different non-dominating subtree, while the ambiguous NP/S NP remains within the same subtree throughout the reanalysis. If true, the time course of processing during the critical region of garden path constructions would only depend on the similarity or dissimilarity of the associated syntactic structures and should not depend on the occurrence frequencies of the associated structures.

However, 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 predictions for those constructions. Since RNN predictions are solely based on the occurrence frequencies in the training data, 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. In fact, Grodner et al. (2003) hypothesized a repair mechanism that could produce surprisal-like effects but with exaggerated magnitudes, which could explain the larger magnitude of the human MV/RR effect compared with the RNN. In particular, 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 would predict exaggerated predictability effects since most predictability effects aside from the probability of the suppressed parse would impact both the initial parse and the reanalysis parse.

4.1 | RNN vs. PCFG language models

Being specifically trained to track syntactic probabilities, PCFG parsers directly represent the difference between the competing syntactic alternatives involved in each garden path effect. In contrast, RNN LMs must infer the probabilities of the syntactic alternatives purely from unannotated lexical sequences and so might be expected to predict less

human-like garden path effects than PCFGs. However, our RNN and PCFG language models made comparable reading time predictions for NP/S and NP/Z constructions; in MV/RR constructions, RNN surprisal predicted the empirical reading times better than PCFG surprisal. While the RNN maintains a representation of the current syntactic state that is less discrete and so is less likely to completely lose one of the relevant parses, the PCFG tracks a finite number of possible parses in a beam and so can completely lose track of very low probability alternatives. However, the PCFG fails to correctly predict *any* MV/RR disambiguation effect even when the analysis is constrained to those cases where the PCFG was able to ultimately determine the correct parse. The PCFG supervised training advantage may be offset by the increased amount of unannotated training data given to the RNN (80 times the data of the PCFG). Indeed, our results could be interpreted as additional evidence that better model accuracy leads to better reading time prediction, in line with findings made on broad-coverage reading time data (Goodkind and Bicknell, 2018).

RNN surprisal was a better predictor of human reading times than was PCFG surprisal, but it is more opaque than a PCFG which outputs a complete description of its final interpretation of each sentence. Therefore, we analyzed the parts-of-speech predicted by the model in the critical region and verified that they were reasonable. 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* "." is not predicted while "," is). These qualitative analyses indicate that the probability model within the RNN do track the correct set of interpretations, which reassures us that the surprisal estimates are valid.

Our present results suggest that there are additional syntactic repair mechanisms which are needed to correctly predict the magnitude of the MV/RR and NP/Z disambiguation response in reading times. These results also suggest that there are different kinds of processes which produce different kinds of garden path responses. Language model improvements or better estimation of spillover effects as recently suggested by Shain and Schuler (2018) could enable model predictions to better anticipate the NP/S reading time responses. The same could be true if the surprisal-to-milliseconds conversion rate we adopted from Smith and Levy (2013) was replaced with a larger conversion rate. However, the models' predictions for NP/Z and MV/RR model sentences are so qualitatively different from the human responses that it seems unlikely that any such changes could account for the mismatch between models and humans. Therefore, it is plausible that NP/S garden path effects could simply be an example of the general effect of predictability which is typically present in reading times, while the NP/Z and MV/RR effects could be caused by a syntactic repair (Sturt, 1997; Sturt et al., 1999) or reprocessing mechanism (Grodner et al., 2003). Alternatively, surprisal may provide a better fit to human reading times when augmented with the noisy channel hypothesis (Levy, 2008; Bicknell and Levy,

2010; Gibson et al., 2013). Either way, our findings indicate that surprisal from current language models is unable to predict the the full range of empirical findings from studies of human processing of temporary ambiguous sentences, and additional modeling assumptions are necessary to address this shortcoming.

ACKNOWLEDGMENTS

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

REFERENCES

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.

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

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.

Francis, W. N. and Kucera, H. (1979) The Brown Corpus: A standard corpus of present-day edited American English.

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.

Futrell, R., Wilcox, E., Morita, T. and Levy, R. (2018) RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Futrell, R., Wilcox, E., Morita, T. and Levy, R. (2019) RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. In *Proceedings of NAACL*.

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.

Goodkind, A. and Bicknell, K. (2018) Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of CMCL*.

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 NAACL*.

Hale, J. (2001) A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*.

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

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.

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. F. (2016) Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 1–30.

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

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

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.

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*.

Nguyen, L., van Schijndel, M. and Schuler, W. (2012) Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of COLING*.

Petrov, S., Barrett, L., Thibaux, R. and Klein, D. (2006) Learning accurate, compact, and interpretable tree annotation. In *Proceedings of ACL*.

Pollard, C. and Sag, I. (1994) *Head-driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

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.

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

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.

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 Schijndel, M., Exley, A. and Schuler, W. (2013) A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, **5**, 522–540.

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.

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.

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

In this paper, we conducted a total of 57 significance tests. We conducted 6 two sample t-tests of whether the spillover-corrected predictions differed from the human reading times (2 models $\times$ 3 constructions). We conducted 6 one-sample t-tests of whether each spillover-corrected model prediction differed from 0 (2 models $\times$ 3 constructions). We conducted 6 one sample t-tests of whether mean non spillover-corrected surprisal differed from 0 (2 models $\times$ 3 constructions). We conducted 6 two sample t-tests of whether mean non spillover-corrected surprisal differed between constructions (2 models $\times$ 3 construction pairings). We conducted 6 two sample t-tests of whether mean spillover-corrected model predictions differed between constructions (2 models $\times$ 3 construction pairings). We conducted 27 two sample t-tests of whether each spillover-corrected model prediction or human response at each word in the critical region differed from each other word in the critical region of that construction ((2 models + 1 human) $\times$ 3 word positions $\times$ 3 constructions). $6 \cdot 5 + 27 = 57$ total significance tests.

1 Supplementary Materials for "Neural network surprisal predicts the existence but not the magnitude of human syntactic disambiguation difficulty"

1.1 Recovery-failure filtered analyses

The PCFG LM fails to recover the correct parse for certain items, which could indicate that the correct parse became so dispreferred that it was no longer considered by the model. Since the surprisal component of a garden path effect is the difference in probability mass between the correct parse and the incorrect distractors, failure to track the correct parse may have a dramatic effect on the ability of an LM to predict the magnitude of the garden path effect. Recovery failure occurred most frequently in MV/RR sentences (53% of the items), followed by NP/S sentences (35% of the items), and least frequently in NP/Z sentences (30% of the items).

This section explores the impact that these difficult parses had on the predictions of each model. We plot the model predictions with all items compared with the model predictions when those difficult items are removed (Figs 1–2). Surprisingly, the effect of filtering out difficult items is not very large.

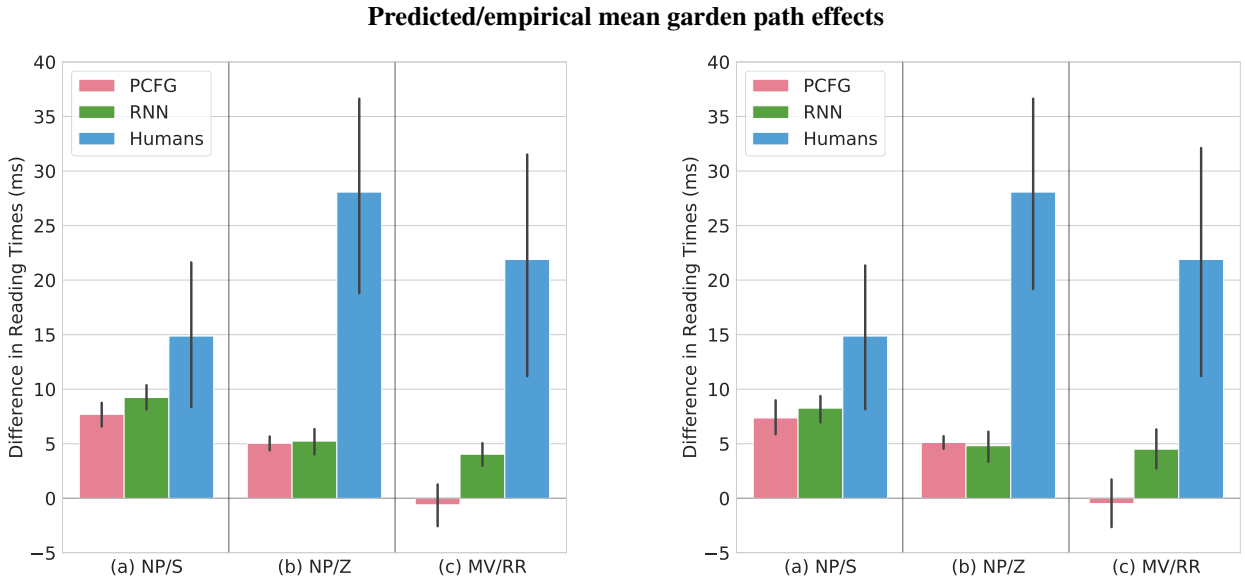

Figure 1: (Left Unfiltered; Right Filtered) Reading time differences predicted by the language models when spillover is accounted for each word. Each subplot shows the disambiguation region of: (Left) ambiguous NP/S sentences compared to matched unambiguous controls (example (4) in the text) (Center) ambiguous NP/Z sentences compared to matched unambiguous controls (example (5) in the text) (Right) ambiguous MV/RR sentences compared to matched unambiguous controls (example (6) in the text). The bar plots show the mean prediction across items; the error bars represent 95% confidence intervals.

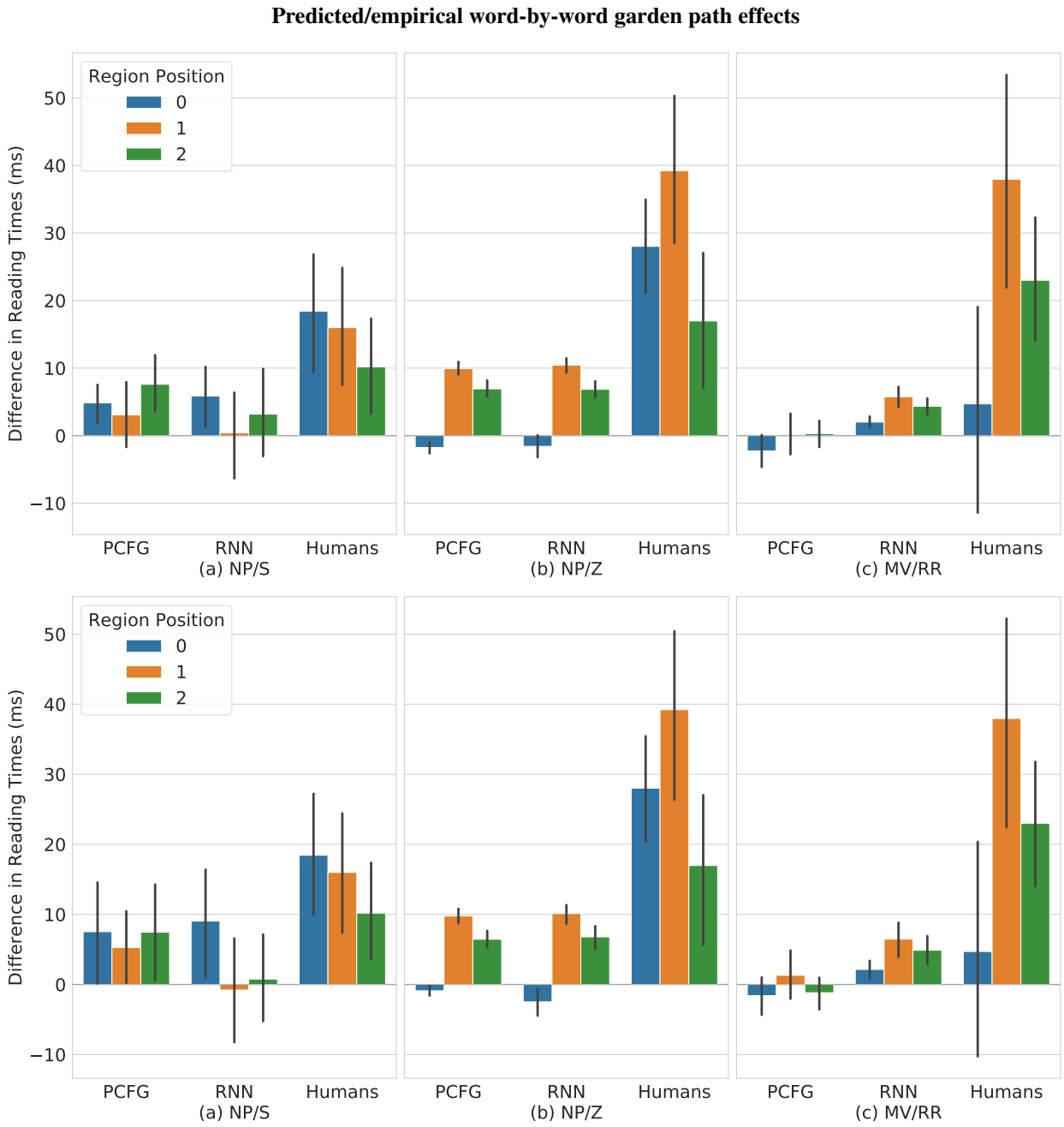

Figure 2: (Top Unfiltered; Bottom Filtered) Reading time differences predicted by the language models when spillover is accounted for each word. Each subplot shows the disambiguation region of: (Left) ambiguous NP/S sentences compared to matched unambiguous controls (example (4) in the text) (Center) ambiguous NP/Z sentences compared to matched unambiguous controls (example (5) in the text) (Right) ambiguous MV/RR sentences compared to matched unambiguous controls (example (6) in the text). The bar plots show the mean predictions across items at word 0, word 1, and word 2 of the critical region; the error bars represent 95% confidence intervals.

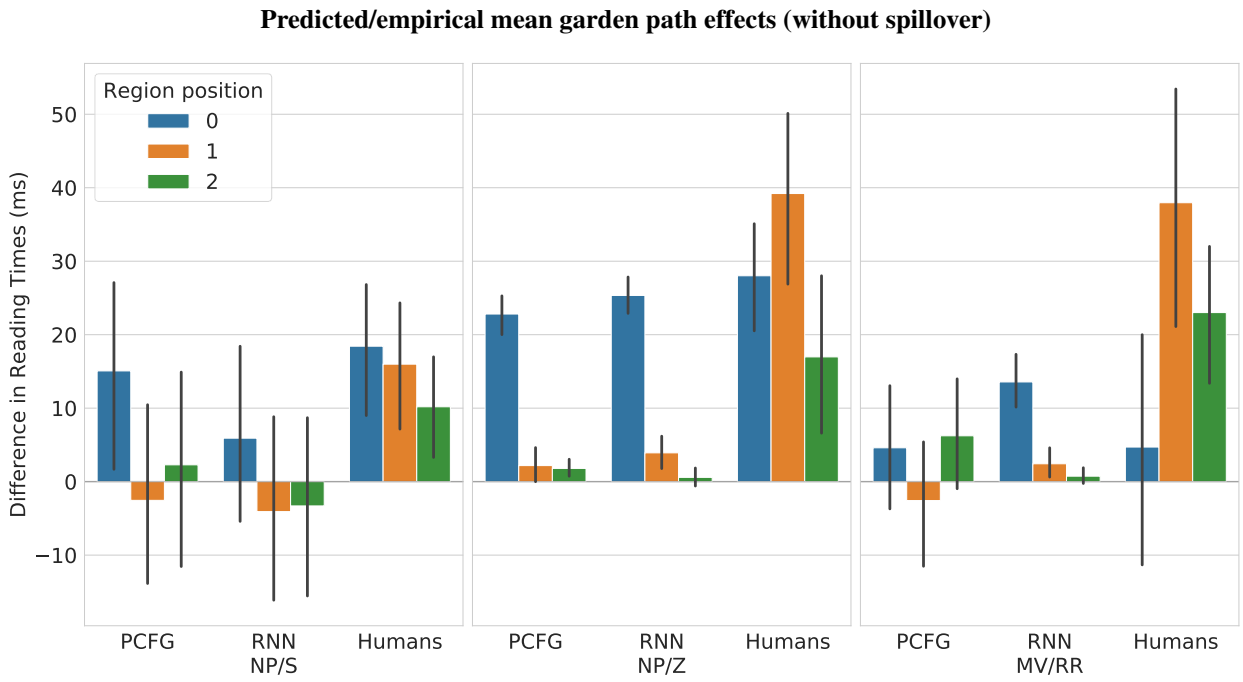

Figure 3: Unfiltered reading time differences predicted by the language models when spillover is not accounted for each word. Instead, surprisal is simply scaled by the 3.75 ms/bit mean surprisal influence observed by Smith and Levy (2013). Each subplot shows the disambiguation region of: (Left) ambiguous NP/S sentences compared to matched unambiguous controls (example (4) in the text) (Center) ambiguous NP/Z sentences compared to matched unambiguous controls (example (5) in the text) (Right) ambiguous MV/RR sentences compared to matched unambiguous controls (example (6) in the text). The bar plots show the mean predictions across items at word 0, word 1, and word 2 of the critical region; the error bars represent 95% confidence intervals.

References

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.