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

Marten van Schijndel PhD[1,*] and 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

[*]Corresponding author: Marten van Schijndel, 203 Morrill Hall, Ithaca, New York 14853, *Keywords:* self-paced reading, garden paths, neural networks, surprisal, information-theory

Affiliation

Abstract

The disambiguation of a syntactically ambiguous sentence in favor of a less preferred 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 initially only maintain a subset of the possible parses of the sentence, and subsequently require time-consuming reanalysis to reconstruct a discarded parse. A more recent proposal argues 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. Since predictability has pervasive effects in reading far beyond garden path sentences, this account, which dispenses with reanalysis mechanisms, is more parsimonious. Crucially, it predicts a linear effect of surprisal: the garden path effect is expected to be proportional to the difference in word surprisal between the ultimately correct and ultimately incorrect interpretations. To test this prediction, we used recurrent neural network language models to estimate word-by-word surprisal for three temporarily ambiguous constructions. We then estimated the slowdown attributed to each bit of surprisal from human self-paced reading times, and used that quantity to predict syntactic disambiguation difficulty. Surprisal successfully predicted the existence of garden path effects, but drastically underpredicted their magnitude, and failed to predict their relative severity across constructions. We conclude that a full explanation of syntactic disambiguation difficulty may require recovery mechanisms beyond predictability.

Keywords: self-paced reading, garden paths, neural networks, surprisal, information-theory

Introduction

Ambiguity is pervasive in human language, and syntactic structure is no exception. In many temporarily ambiguous sentences, where 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.

After they have read the first few words of (1), readers tend to prefer the interpretation in which the girl phoned the instructor; in other words, they prefer to parse *the instructor* as the direct object of *phoned*. When the reader reaches the subsequent verb *was*, it becomes clear that this initially preferred interpretation leaves no viable subject for this verb. Empirically, reading times at the disambiguating region *was very upset* are elevated compared to those measured on the same words when they are 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*: when the comma is present, 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 in 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 the sentence, readers only consider one of the possible partial parses of the sentence (Frazier and Fodor, 1978; Pritchett, 1988), or only consider a small number of possible parses (Gibson, 1991; Jurafsky, 1996). In those theories, processing difficulty in the critical region arises as a consequence of the reanalysis required to reconstruct a parse that was initially entertained and then discarded, or not considered in the first place, but that later turned out to be correct (Pritchett, 1988; Gorrell, 1995; Sturt and Crocker, 1996; Sturt, 1997; Bader, 1998). We refer to these theories as *two-stage accounts*.

In contrast with such two-stage accounts, some recent 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 one mechanism is simpler than a theory based on two. For example, if word predictability—an independently motivated predictor of reading times (Ehrlich and Rayner, 1981; Rayner and Well, 1996; Smith and Levy, 2013)—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 in prior work has accounted for syntactic disambiguation difficulty more successfully than entropy reduction (Linzen and Jaeger, 2016). Under the surprisal account, syntactic disambiguation difficulty emerges 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). In the case of 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, which, in turn, is due to the fact that it is only consistent with a low-probability parse (Hale, 2001; Levy, 2013).

Computational simulations have demonstrated that the words of the critical region are indeed less predictable in temporarily ambiguous sentences than in unambiguous controls. Levy (2013), for example, showed this to be the case 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). While such simulations are *consistent* with the predictions of surprisal theory, we argue, the conclusion that garden path effects can be *reduced* to predictability is premature, for two reasons. First, if surprisal alone is expected to explain 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, in the following sentence, as in (1) above, a noun phrase (here *the contract*) is initially likely to be interpreted as a direct object; unlike in (1), however, the disambiguating word (*would*) signals that 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 so-called NP/S sentences such as (3) to NP/Z sentences such as (1), the empirical slowdown measured at the disambiguation point is more moderate 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 the disambiguating word *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 one-stage models face 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 linear relationship between surprisal and reading time—or, equivalently, the same logarithmic relationship between predictability and reading time—should hold 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—resulting in 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 $70/4 = 17.5$ 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 adopt the taxonomy of cognitive model predictions proposed by Padó et al. (2009), these

challenges to surprisal theory arise from the fact that the theory not only makes qualitative predictions as to the *existence* of a processing difficulty, but 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 or the entropy reduction hypothesis compared to verbal models, which make much weaker predictions; but, we argue, these stronger predictions invite a more detailed empirical assessment than has been attempted in the past. In this work, we investigate the ability of quantitative single-stage models to predict, in each of these three senses, the garden path effects observed in human self-paced reading studies.

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 the proportion of participants who completed the fragment with *was*. While this method can distinguish highly predictable words (e.g., $P(w|\text{context}) = 0.8$) from moderately predictable words (e.g., $P(w|\text{context}) = 0.1$), 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 performing the cloze task occasionally considered the dispreferred parse before the disambiguation point, millions of participants may be required to accurately estimate the very low probabilities that, according to surprisal theory, likely characterize the disambiguating words in garden path sentences.

An alternative approach to estimating the predictability of words relies on probabilistic *language models*, computational systems 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}, which, as discussed in the introduction, may be required to derive a 70 ms empirical effect from surprisal theory.

Probabilistic language models can be based on a range of computational architectures. Many of the words in typical sentences can be predicted well from local context using n-gram models,

which are based on tabulating the frequency of 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 much larger contexts. 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 explicitly designed to construct syntactic parses, and are not provided any syntactic annotations during training, recent empirical studies have shown that the probability distributions defined by those models reflect sensitivity to a range 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.

Overview of experiments

To test the surprisal account of garden path effects, we use surprisal estimates derived from RNN language models to simulate the results of publicly available self-paced reading data. 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 the Materials section. To estimate the overall correlation between language model surprisal and reading times, we use reading times on filler sentences that do not contain these three types of temporary syntactic ambiguity; we then apply the same correlation coefficient to predict reading times from model-derived surprisal for garden path sentences. In calculating the slowdown that can be attributed to a particular unpredictable word, we pay careful attention to the possibility that self-paced reading time on a given word reflects processing difficulty on an earlier word (spillover).

To anticipate our results, when averaged over the disambiguating region, RNN surprisal correctly predicted a slowdown in the disambiguating region of ambiguous sentences, compared to unambiguous controls, in all three constructions; in other words, 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 empirically observed slowdown in all three constructions. The discrepancy varied across constructions: it was small in NP/S, moderate in MV/RR, and very large in NP/Z. Surprisal predicted numerically larger disambiguation difficulty in NP/S than NP/Z sentences, the opposite pattern from humans. Finally, the detailed word-by-word contour of the garden path effect over the disambiguating region was not well-predicted by surprisal, even at the qualitative level of explanation. 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.

Methods

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 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.

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 verb *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, with the interpretation that the soldiers were the ones warning about the dangers, or as the verb of a reduced relative (RR) clause, with the interpretation that 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. In the matched unambiguous version of each construction, these words are, of course, not disambiguating; to refer to these three words in both contexts we will also use the term "critical region".

Self-paced reading measurements

We focus our modeling efforts on reading times measured using the moving-window self-paced reading paradigm (Just et al., 1982). In this 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. This paradigm rests on the assumption that processing difficulty on a word can cause participants to delay advancing to the next word of the sentence, although in practice any such delays are often observed on subsequent words ("spillover"; see below).

We use the publicly available self-paced reading times reported by Prasad and Linzen (2019a)

and Prasad and Linzen (2019b). Prasad and Linzen (2019b) had online participants recruited on Amazon Mechanical Turk (224 subjects after standard exclusions) read sentences with NP/S and NP/Z ambiguities. The materials were adopted from Grodner et al. (2003); the ambiguous noun phrase was always a plausible object of the verb (cf. Garnsey et al. 1997). Prasad and Linzen (2019b) 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. Importantly, participants in both studies also read filler sentences, with a variety of unambiguous syntactic structures, as is standard in self-paced reading studies; we use these filler items 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 if, as is often the case, publication is contingent on obtaining a statistically significant result (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; a point we will return to in the General Discussion.

Language models

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

[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`

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; Frank and Hoeks, 2019), among other syntactic phenomena.

Since Wikipedia sentences may be longer and more complex than the sentences that make up the bulk of the linguistic experience of participants in the reading studies we model, an RNN language model trained on Wikipedia may assign unrealistically high probability to complex constructions such as the ones we investigate, leading our model to systematically underpredict garden path effects. To address this concern, we trained another RNN language model on a soap opera dialog corpus (Davies, 2011), using similar parameters to those used to train Wiki RNN;[2] we refer to this model as Soap RNN. The average sentence length in the soap opera training corpus is nine words, much shorter than the average sentence length of the Wikipedia training corpus (27 words). Visual inspection of this corpus suggests that the syntactic structures it includes tend to be much simpler than those that are typical in Wikipedia.

Accounting for spillover

The surprisal of a word affects self-paced 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 not only by the surprisal of the words of the critical region, but also by the surprisal of material preceding

[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.

the critical region. For concreteness, consider the MV/RR sentence (7a):

(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". Since disambiguation occurs entirely at the first word of the critical region (*conducted* in the above example)—subsequent words of the critical region do not provide any additional information about the relevant parsing decision—surprisal on the second and third word of the critical region is identical across conditions. 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 the difference in the complex pattern of spillover influence due to surprisal. These considerations suggest that a spillover-adjusted linking function is essential for predicting reading times from surprisal. We describe such a linking function in the following section.

Estimating the quantitative effect of surprisal on reading times

Smith and Levy (2013) show that, all else being equal, there is a linear relationship between the surprisal of a word and reading times on that word and the following ones. The coefficient of this linear relationship varies depending on the distance between the word whose surprisal is considered and the word on which reading times are measured. We apply the procedure described by Smith and Levy to compute these coefficients, which we refer to as *conversion factors*, from the reading times for the **filler sentences** from Prasad and Linzen (2019b), sentences that do not include examples of

the three types of temporary ambiguity in question. To foreshadow the conclusions of the analysis we present in this section, the total surprisal-to-RT conversion factor, when summed across the word whose surprisal is considered and the three subsequent words, was approximately 2 ms/bit for both of our models.

To estimate the conversion factors, 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_i) and the preceding three words (w_{i-3}, w_{i-2} and w_{i-1}): surprisal (S_{i-3}, S_{i-2}, S_{i-1}, S_i), entropy ($H_{i-3},\ldots,H_i$), entropy reduction ($\Delta H_{i-3},\ldots,\Delta H_i$), word frequency ($f_{i-3},\ldots,f_i$), word length ($l_{i-3},\ldots,l_i$), and the position of the word in the sentence (p_i). Entropy and entropy reduction were computed based on the probability distribution over the vocabulary defined by 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 (for discussion, see Linzen and Jaeger 2016), as the latter is intractable to compute for RNN language models. 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 $f_{i-1} : l_{i-1}$ but not $f_{i-1} : l_{i-3}$). Finally, we included by-participant random intercepts. Formally, our model was as follows:

$$\begin{aligned}\text{rt} \sim & S_i + S_{i-1} + S_{i-2} + S_{i-3} + H_i + H_{i-1} + H_{i-2} + H_{i-3} + \Delta H_i + \Delta H_{i-1} + \Delta H_{i-2} + \Delta H_{i-3} + \\ & p_i + l_i * f_i + l_{i-1} * f_{i-1} + l_{i-2} * f_{i-2} + l_{i-3} * f_{i-3} + (1 \mid \text{subject}) \end{aligned} \quad (1)$$

where the notation $x * y$ indicates that x, y and their interaction $x : y$ were all included in the model. We reiterate that this method to compute the conversion factor relies on the theoretical assumption, confirmed by Smith and Levy (2013), that surprisal has a linear effect on reading times. The entropy reduction hypothesis similarly predicts a linear effect of entropy reduction on reading times; there is some empirical evidence of the effectiveness of entropy reduction in predicting reading times (Frank 2010; Linzen and Jaeger 2016; Lowder et al. 2018; for a less positive conclusion, see Aurnhammer and Frank 2019). The remaining predictors included in the regression 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 factor from each

of the complexity metrics to milliseconds of reading time. We only retain regression coefficients that were significantly different from zero at the $p < 0.01$ level. In the case of surprisal, for example, if the coefficients for S_{i-1}, S_{i-2}, and S_{i-3} were significant, we concluded that these surprisal values robustly mapped onto reading times for filler sentences, and that they were therefore predicted to map robustly onto the reading times for garden path sentences, if those were driven by surprisal.

In what follows, we refer to the conversion factors as δ_{-3}, δ_{-2}, δ_{-1} and δ_0. In the case of surprisal, those would correspond to S_{i-3}, S_{i-2}, S_{i-1} and S_i, respectively. Our analysis did not reveal a significant effect of the surprisal of a word on its own reading time ($\delta_0 S_i$). 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-adjusted surprisal effect $\hat{S}_i$ on the i-th word of the sentence as follows:

$$\hat{S}_i = \delta_{-3} S_{i-3} + \delta_{-2} S_{i-2} + \delta_{-1} S_{i-1} \tag{2}$$

For the Wiki RNN language model, our estimates of the individual spillover conversion factors for surprisal were $\delta_{-1} = 1.1$ ms/bit, $\delta_{-2} = 0.37$ ms/bit, and $\delta_{-3} = 0.39$ ms/bit (the full set of conversion factors, for the two language models and three complexity metrics, is given in Table 1). These conversion factors 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.

Analyses

Overview

Before we discuss our analyses, we briefly reiterate the logic behind them. Recall that surprisal theory assumes that every 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

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 factors for each information-theoretic measure for each RNN. We only report conversion factors 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.*

entirely by the conditional probability of the disambiguating words, this surprisal-to-RT conversion should be sufficient to fully explain the magnitude of the garden path effect measured in a self-paced reading study.

We report three analyses that rely on this logic. Analysis 1 follows the traditional analysis approach in the human behavioral literature and aggregates 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 spillover-adjusted linking function, correctly identifies the precise locus of processing difficulty in each type of ambiguity. As we will see, both Analysis 1 and Analysis 2 identify significant discrepancies between models and humans. Analysis 3 extends the methodology we introduce in Analysis 1 to next-word entropy as well as entropy reduction computed from next-word entropy.

We then report two control analyses. Analysis 4 tests the hypothesis that the failure of RNN surprisal to predict the magnitude of the garden path effect is due to a floor effect, where the language models simply never assign low enough probabilities in any context, either in garden path sentences

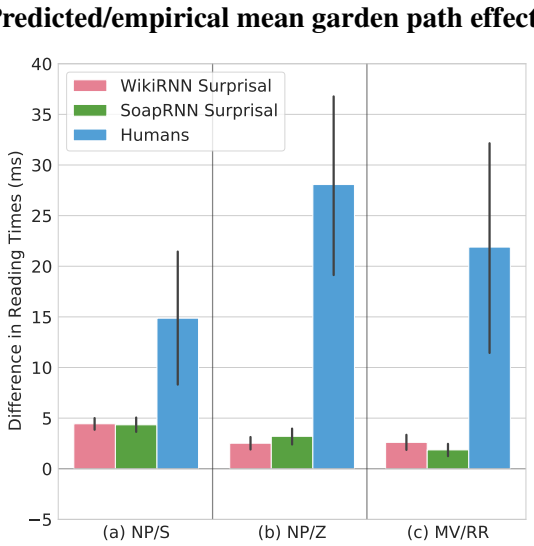

Figure 1. Difference in reading times between ambiguous and unambiguous sentences, averaged over the three words of the critical region, as predicted by the Wiki RNN language model (in pink) and the Soap RNN language model (in green), compared to empirical reading times on the region (in blue). Each 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 bootstrapped 95% confidence intervals.

or elsewhere; we do not find support for this hypothesis. Finally, Analysis 5 shows that the 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 RNNs' failure to analyze the syntactic structure of the sentence.

Analysis 1: Aggregating predicted and empirical reading times over the critical region

We first report an analysis that aggregates reading times and model predictions over the three words of the critical region, following standard practice in the sentence processing literature. Using the approach described in the methods section, 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 (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 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 difference between predicted and empirical RTs was highly significant for both Wiki RNN (NP/S: $p = 0.005$; NP/Z: $p < 0.001$; MV/RR: $p < 0.001$) and Soap RNN (NP/S: $p = 0.006$; NP/Z: $p < 0.001$; MV/RR: $p < 0.001$).[3]

The conclusions of Analysis 1 are straightforward. If, as argued by surprisal theory, the relationship between surprisal and reading times is linear, and surprisal accounts for the entire processing difficulty observed in the disambiguating region of garden path sentences, then a conversion factor 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 suggest that that is not the case.

Analysis 2: Predicting word-by-word reading times

Analysis 1 examined the garden path effect averaged over the critical region, following standard practice in the analysis of studies of human sentence processing using self-paced reading. To obtain a more fine-grained picture of the models' predictions, we next examined 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). As before, statistical significance was assessed using t-tests for each sentence position, paired across the ambiguous and unambiguous version of each item.[4]

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 of the MV/RR construction, where neither humans nor RNNs

[3]Throughout this paper, we report raw p values, not corrected for multiple comparisons. In each section, we explicitly list our statistical tests to enable post-hoc corrections for multiple comparison. In general, however, the results we report are robust enough that most p values survive correction for multiple corrections. In Analysis 1, we conducted 12 t-tests of whether the mean spillover-adjusted predictions differed from the mean human reading times (paired by item) or from 0 (1-sample): 2 models $\times$ 3 constructions $\times$ 2 comparisons.

[4]In Analysis 2, for each of the two spillover-adjusted model predictions and the human responses, we conducted 18 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 constructions $\times$ 6 comparisons.

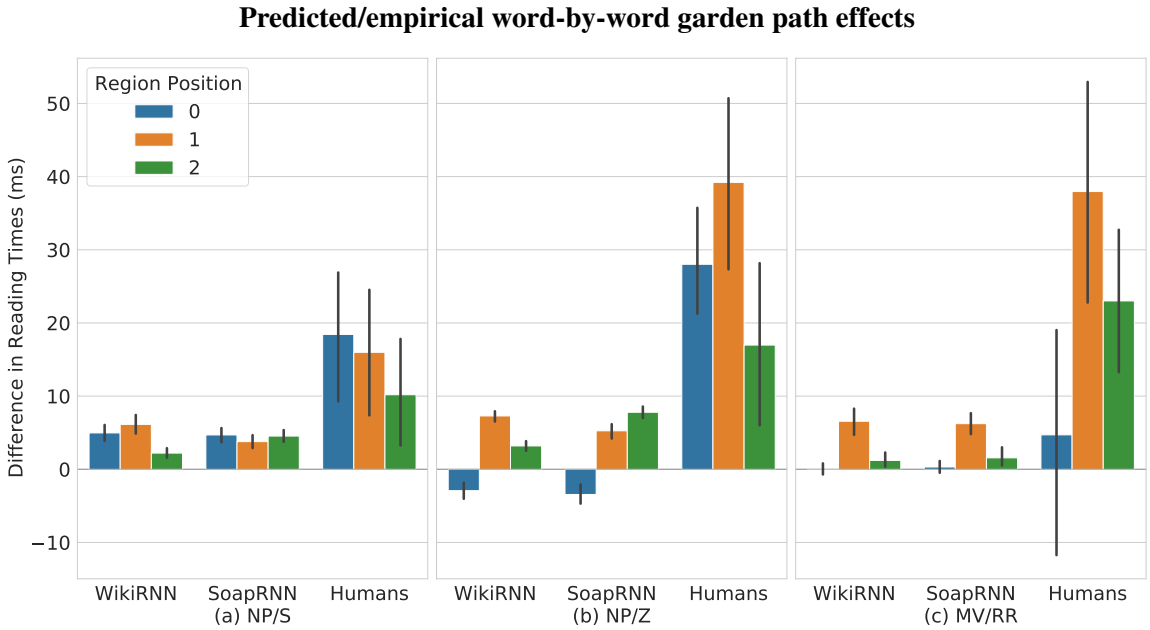

Figure 2. Differences in word-by-word reading times between ambiguous and unambiguous sentences on the first, second and third word of the disambiguating region, as predicted by the language models, compared to empirical reading times. The subplot shows the disambiguation region of: (a) ambiguous NP/S 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 bootstrapped 95% confidence intervals.

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-adjusted predicted RTs and empirical RTs to validate our adjustment for spillover. In the remainder of this section, we discuss the detailed empirical and predicted pattern for each construction.

NP/S. In this construction, the empirical effect was spread over the entire critical region, with no significant differences between any two points in the region (all $p > 0.05$), though there was a numerical decrease over the course of the critical region. Qualitatively, the predictions derived from Wiki RNN matched the empirical pattern, with a decrease over the critical region; unlike in the empirical reading times, this decrease in reading times reached significance in Wiki RNN's predictions ($p < 0.001$). By contrast, 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.

NP/Z. In NP/Z sentences, the first and second words of the critical region carried the bulk of the empirical garden path effect: the third word of the critical region showed 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 exhibited a large NP/Z garden path effect in both the first and second word of the region.

MV/RR. The second and third words of the critical region carried most of the human garden path effect in the MV/RR ambiguity, 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 was numerically reduced at the third word, but did not reach statistical significance. Further, the models correctly predicted that the first word in the region should not exhibit an appreciable garden path effect. In summary, 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 ones.

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 made by 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 question, 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 the MV/RR ambiguity than the NP/S ambiguity. This discrepancy suggests that humans process these two constructions in different ways.

It is possible, of course, that some of the discrepancy in magnitude between the empirical and predicted garden path effects arises from an incorrect estimate of the conversion factor between bits of surprisal and milliseconds of reading times. Crucially, however, the fact that this discrepancy differs in magnitude 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.

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;

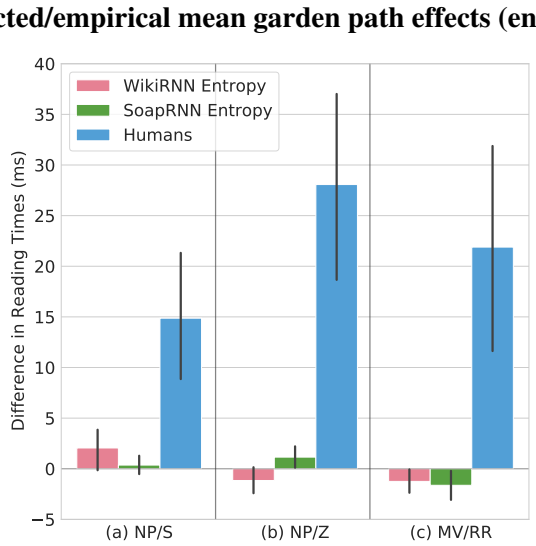

Figure 3. Difference in reading time between ambiguous and umambiguous sentences, averaged over the three words of the critical region, as predicted by the entropy of the Wiki RNN language model (in pink) and the Soap RNN language model (in green), compared to empirical reading times on the region (in 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 bootstrapped 95% confidence intervals.

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 factors 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).[5]

We found that entropy and entropy reduction were much poorer predictors of human garden path effects than surprisal (Figs. 3 and 4). In fact, in most cases these measures predicted no effect

[5]In Analysis 3, we conducted 12 t-tests of whether the mean spillover-adjusted predictions differed from the mean human reading times (paired by item) or from 0 (1-sample): 2 models $\times$ 3 constructions $\times$ 2 comparisons.

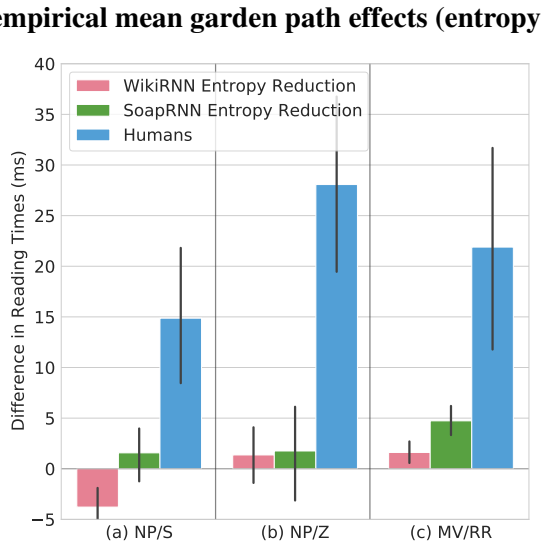

Figure 4. Difference in reading time between ambiguous and umambiguous sentences, averaged over the three words of the critical region, as predicted by the entropy reduction of the Wiki RNN language model (in pink) and the Soap RNN language model (in green), compared to empirical reading times on the 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. Error bars represent bootstrapped 95% confidence intervals.

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 unable to test that hypothesis as we are not aware of methods that can estimate full-sentence entropy from RNN language models. However, our results are consistent with those of Linzen and Jaeger (2016), who did derive full-sentence entropy from a grammar-based language model, and found that entropy reduction computed in this way did not predict a garden path effect in the correct

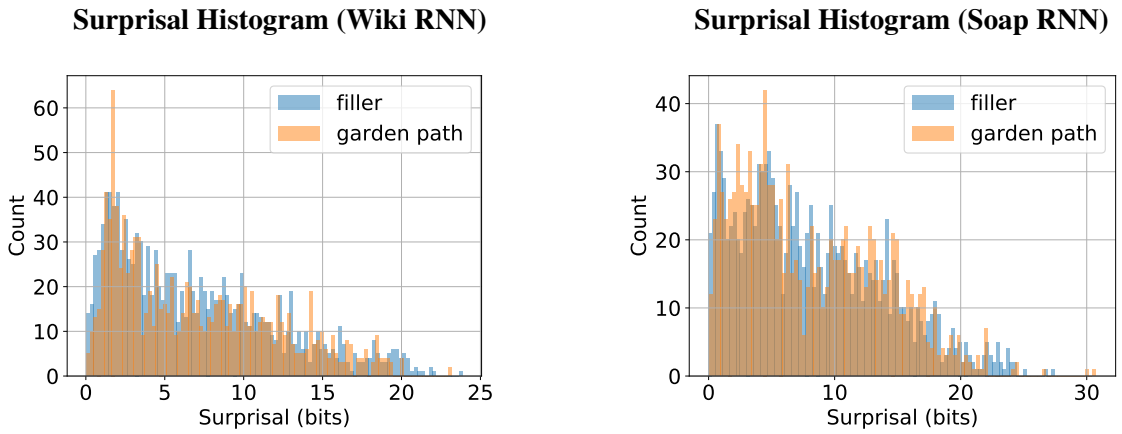

Figure 5. Count histograms of binned (N=100) surprisal values over all tokens in filler sentences (in blue) and garden path sentences (in orange), as estimated from the Wiki RNN (Left) and Soap RNN (Right).

direction.

Analysis 4: Can RNN language models assign sufficiently low probabilities?

For the surprisal hypothesis to derive the full magnitude of garden path effects, surprisal in garden path sentences needs to be fairly large. The human effect size for NP/Z sentences reported by Prasad and Linzen (2019b), for example, was approximately 28 ms; with a conversation factor of 2 ms per bit of surprisal this entails that the surprisal of the first word of the ambiguous region needs to be 14 bits higher in ambiguous than unambiguous sentences. If we assume that the surprisal of this word in unambiguous items is around 3–5 bits, then its surprisal in ambiguous sentences needs to be around 18 bits (equivalent to a probability of about 0.000004), and, assuming some variability across sentences, even larger than that for some of the items. Could our underestimates of the magnitude of the effect be due to a ceiling on the surprisal values that our language models can produce, regardless of the context? To determine whether that is the case, we binned the surprisal values for each language model for each input token in the Prasad and Linzen (2019b) dataset, separately for filler sentences and for NP/Z and NP/S sentences. We used 100 bins. We omitted from the analysis the first token of each sentence, which was very often the word "The"; this was done to avoid distorting the histograms with a large number of identical surprisal values.

As shown in Figure 5, there was not a sharp ceiling for surprisal values, which we might expect

if models were simply unable to capture very infrequent events; in fact, the surprisal values they predicted, especially for filler sentences, exceeded 18 bits with some regularity. We cannot rule out the possibility that the models' probabilities for rare events are more poorly calibrated than those assigned to common events. At the same time, we emphasize that the probabilities assigned by the models would need to be systematically biased in the same direction, and by orders of magnitude, for the surprisal hypothesis to remain viable.

Analysis 5: Do RNN language models make appropriate syntactic predictions?

Could the models' inability to predict the magnitude of the human garden path effect be due to a broad failure to take the syntactic structure of temporarily ambiguous sentences into account when making word predictions? Such a lack of sensitivity to syntactic structure would entail that the particular language models we used in this paper cannot be used to address the viability of the surprisal hypothesis. While work that suggests that the predictions of RNN language models are in fact sensitive to various syntactic constraints provides reason for optimism (Linzen et al., 2016; Wilcox et al., 2018; Futrell et al., 2019), the goal of the current section is to explore the validity of this concern for the particular constructions and items used in this study.

Since the predictions made by the two language models were qualitatively similar to one another, we focus our analysis in this section on Wiki RNN. As a window into this model'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 Wikipedia corpus used by Linzen et al. (2016). 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 resulting probability distribution over upcoming parts-of-speech in ambiguous sentences, as compared 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 verbs, 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

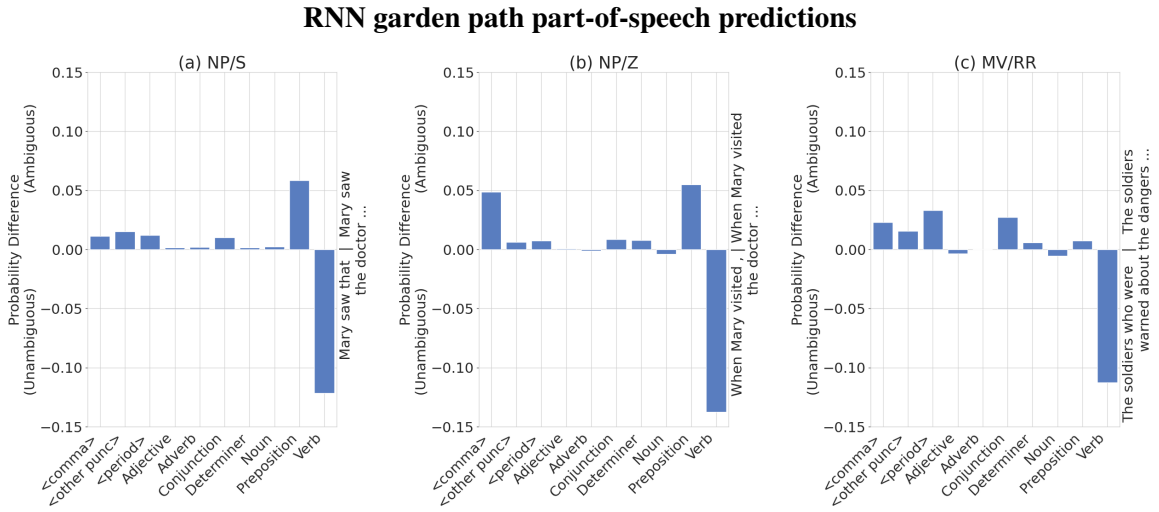

Figure 6. Part-of-speech predictions of the recurrent neural network language model on the first word of the critical region of unambiguous sentences, subtracted from the predictions on the same word in their ambiguous counterparts, for (a) NP/S sentences, (b) NP/Z sentences, and (c) MV/RR sentences. Positive values indicate that the part-of-speech in question is more likely in ambiguous sentences, and negative values indicate that it is more likely in unambiguous sentences.

particular, in ambiguous NP/S sentences (Fig. 6a), the model generally encoded the expectation that a prepositional phrase should appear next (for example, *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. 6b), 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 (e.g., a comma), both of which are continuations consistent with the ultimately incorrect "NP" parse, where *doctor* is the object of *visited*. Lastly, in MV/RR ambiguous sentences (Fig. 6c), the model predicted a punctuation mark would come next, most likely a period (*The soldiers warned about the dangers* .), consistent with the preferred but ultimately incorrect main verb (MV) parse.

Overall, Wiki RNN's predictions reflect sensitivity to the syntactic structure of the temporarily ambiguous sentences used in our experiments; 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.

General Discussion

Garden path sentences are temporarily ambiguous sentences that are eventually disambiguated in favor of the initially dispreferred parse. In those sentences, reading times at the disambiguation point are elevated compared to matched unambiguous control sentences; this relative slowdown is referred to as a garden path effect. A number of accounts have attributed this slowdown to the processing cost incurred by 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-stage 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 a conversion factor quantifying the effect of predictability on reading times in filler sentences, which did not include garden path constructions. In estimating the conversion factor, we took into account 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, trained on either Wikipedia articles or soap opera dialogues. Finally, we multiplied these surprisal estimates by the conversion factor to generate predicted reading times for the disambiguating region of garden path sentences, which we then compared to empirical reading times from human experiments.

While the language models indeed predicted higher surprisal in the critical region of temporarily ambiguous sentences compared to control sentences (in line with Hale 2001; Levy 2013; Futrell

et al. 2019), the difference in surprisal between the ambiguous and unambiguous versions of each sentence 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 linking functions of surprisal to human behavior have no hope of deriving the human pattern, even if the true conversion factor between surprisal and RTs is very different from the one we estimated.

Finally, we reported two control analyses. First, Analysis 4 inspected the overall distribution of surprisal values assigned to each of the words in the filler, NP/S, and NP/Z sentences, both inside and outside the critical region, and confirmed that RNN language models regularly assign very low probabilities, in the range required to explain garden path effects using surprisal. This finding both justified our use of broad-coverage language models, as opposed to cloze responses—surprisal values of 20 bits or more would be impossible to elicit reliably using the cloze procedure—and indicated that the failure of the models to correctly predict garden path magnitudes was not driven by the RNN language models' inability to assign low enough probabilities to words in general.

Second, in Analysis 5 we verified that the models' syntactic predictions for temporarily ambiguous sentences were consistent with the structure of those sentences. Unlike grammar-based language models, which make available interpretable representations of the structures considered at each point in the sentence, RNNs only output lexical predictions. To address this issue, we inferred the models' syntactic predictions from probabilities aggregated over parts-of-speech. In unambiguous conditions, our RNN language model made the correct prediction of a verb, and in the ambiguous conditions it made alternative predictions consistent with the preferred parse (e.g., given the context *When Mary visited the doctor*, a period indicating the end of the sentence was not assigned a significant probability, while a comma was, appropriately). This analysis supports the hypothesis that the probability distribution defined by the RNN tracks the expected set of syntactic parses.

Word-by-word reading patterns

Traditional analyses target mean reading times in the critical region. While we report an analysis that follows this approach (Analysis 1), we also explored empirical and predicted word-by-word reading patterns throughout the critical region of each of the garden path constructions (in Analysis 2). This second analysis revealed that the empirical 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, when combined with a spillover-adjusted linking function, 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 fact that the contour of the human effect differs by 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 from the subtree representing the subordinate clause (*Even though the girl phoned the instructor...*) to a new subtree, the main clause subtree (*the instructor was very upset*). Since this new subtree is not dominated by the subtree that contained the NP before the transformation, reanalysis difficulty is greater. In the NP/S ambiguity, by contrast, the initially ambiguous NP remains within the same subtree—the main clause verb phrase—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.

The word-by-word human garden path effects in NP/S and MV/RR constructions followed a similar time course to the RNNs' 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 may be related to syntactic co-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). They 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.

Relationship to other sources of processing difficulty

Reading behavior is affected by a range of factors other than surprisal. Those include 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 argued that surprisal needs to be supplemented 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 are unfamiliar with any such proposals, and believe that this possibility is unlikely: factors such as word length or memory retrieval interference are in all likelihood perfectly matched across the ambiguous and unambiguous versions of each type of garden path construction.

Factors other than surprisal do affect the processing of most words in filler sentences, which we used to estimate the conversion factor between surprisal and reading times. We only controlled for one of them (word length) when we estimated the conversion factor, and controlling for additional variables may lead to even more accurate conversion factors. In any case, we do not believe that our conclusions strongly depend on the precision of our estimate of the conversion factor: in fact,

in early analyses not included in the current paper (van Schijndel and Linzen, 2018), we found that surprisal substantially underpredicted garden path effects even when the conversion factor was double the one we used in the current paper.

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). This factor could vary systematically between the ambiguous and unambiguous version of each construction, and across garden path constructions. To address such potential plausibility confounds, previous studies have supplemented language models with explicit models of semantic fit (Padó et al., 2009). We believe this issue represented a greater cause for concern in earlier studies, which computed surprisal using probabilistic context-free grammar models trained on small corpora (approximately one million words). Such language models, while appropriately capturing the syntactic distinctions across conditions, may indeed fail to adequately capture semantic plausibility constraints. 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 and pragmatic generalizations through their distributed representations of words (e.g., Mikolov et al., 2013; Levy and Goldberg, 2014; Schuster et al., 2020), which were unavailable to earlier grammar-based language models. Particularly pertinent to the current work is the study by Frank and Hoeks (2019), who showed that the strength of the garden path effect predicted by RNN language models is modulated by semantic plausibility. Overall, 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).

Converging evidence for two-stage accounts

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 the dissociation observed in the event related potential (ERP) literature between the N400 component, which is sensitive to word predictabil-

ity (Van Petten and Luka, 2012; Frank et al., 2015), and the P600 component, which, while not straightforwardly related to word predictability, is strongly modulated by disambiguation in favor of the dispreferred parse in garden path sentences (Osterhout et al., 1994). Evidence for a dissociation between predictability and reanalysis difficulty from the eye-tracking-while-reading paradigm is 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). As such, its prediction—that less predictable words should be read more slowly—can arise from the aggregate 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 constructions that give rise to the dissociation between N400 and P600. This arguably provides converging support for the existence of a second-stage reanalysis mechanism, which is indexed by the P600, and causes a slowdown in reading that is significantly more severe than predicted by surprisal.

A failure of the surprisal hypothesis or a failure of our language models?

The surprisal hypothesis can only be tested given a particular model that assigns predictability values to individual words. In this paper, we have used for this purpose two RNN language models, trained on two different corpora (Wikipedia and soap opera dialogues). The two models yielded largely converging results: surprisal estimates for the disambiguating word in garden path sentences were insufficient to explain the magnitude of the human garden path effect. We have argued that RNN language models, and in particular those based on the LSTM architecture used by Gulordava et al. (2018), are appropriate for testing the surprisal hypothesis: they are sensitive to syntactic constraints in general (Wilcox et al., 2018; Futrell et al., 2019), and, as we have shown in Analysis 5, make predictions that are qualitatively consistent with the correct analyses of the particular temporary syntactic ambiguities we investigate.

It is certainly possible that a different language model could match the human reading pattern.

This would require surprisal estimates that are substantially higher across the board than those of the models we tested, and, unlike our RNN language models, significantly higher for NP/Z than NP/S constructions. Such differences in language model behavior could arise from alternative architectures, such as a Recurrent Neural Network Grammar, which simultaneously parses the sentence and predicts the next word (Dyer et al., 2016; Wilcox et al., 2019); an RNN trained to jointly predict the next word and a syntactic property of the current word (Enguehard et al., 2017); or architectures that differ from RNNs in their inductive biases in ways that are not explicitly informed by syntactic structure, such as Transformers (Vaswani et al., 2017; Hu et al., 2020; Merkx and Frank, 2020). A closer match between language model and human predictions could also arise from a different training corpus: a text corpus that matched the participants' linguistic experience more closely—for example, one that included a mix of dialogues, child-directed speech, newspaper text, and social media posts—or even a multimodal corpus.

In light of the large space of possible language models, it may be difficult to definitively falsify the surprisal hypothesis. To give an extreme example, one can imagine a modification of one of our RNN language models that explicitly detects each of the three types of temporary ambiguities, and divides the probabilities of words in the disambiguating region by construction-specific factors, such that the resulting surprisal values fit the human results perfectly. In future work, such circularity should be avoided by selecting a language model based on external criteria, such as perplexity (Goodkind and Bicknell, 2018), or the model's generalization abilities in other syntactic contexts (Hu et al., 2020).

The surprisal conversion factor

The analyses reported in this paper were based on the assumption of a linear effect of surprisal on reading time (Hale, 2001). We referred to the slowdown in milliseconds that can be attributed to each bit of surprisal as the *conversion factor*. We conducted all of our analyses at the group level: we estimated a single conversion factor for all participants, based on reading time measurements from filler items, and fit it to the average garden path effect across critical items and participants. This group-level analysis is a simplification, of course. In future studies, more precise analyses might estimate a separate conversion factor for each subject, or interpolate between subject-specific and group-level conversion factor using mixed-effects models. Likewise, future analyses could take

into account any across-item variability in the strength of the garden path effects, and use linking functions based on more sophisticated models of spillover (Shain and Schuler, 2018).

Our estimate of the conversion factor for our data was approximately 2 ms/bit. This contrasts with the 4 ms/bit conversion factor estimated by Smith and Levy (2013). It is likely that the main cause for this discrepancy is the subject population and experimental procedure: the self-paced reading times reported by Smith and Levy (2013) were obtained from undergraduate students who performed the experiment in the lab, whereas our participants were recruited on crowdsourcing platforms and performed the experiment online. Self-paced reading participants recruited on Mechanical Turk read much faster than in-lab participants; Enochson and Culbertson (2015) report an average difference of 180 ms per word between in-lab and online participants. The average garden path effect measured in the data we model is qualitatively consistent with this discrepancy: for NP/Z, for example, the garden path effect for Prasad and Linzen's online participants was 28 ms, compared to 70 ms in the in-lab study of Grodner et al. (2003). Another factor that may have contributed to the difference between our conversion factor and that of Smith and Levy (2013) is the language model used to derive surprisal estimates: Smith and Levy (2013) used a trigram model, which in general produces less accurate probability estimates than the RNN language models we used (Gulordava et al., 2018). It is unclear, however, whether we expect trigram surprisal to be systematically lower than RNN surprisal in filler sentences, and especially to an extent that could result in a substantially higher conversion factor.

We note that the qualitative conclusions of the present work do not strongly depend on the precise conversion factor we used, in two respects. First, in earlier work (van Schijndel and Linzen, 2018), we found that surprisal substantially underestimated the empirical garden path effects even when we used the higher 4 ms/bit conversion factor derived from Smith and Levy (2013). Second, because language model surprisal was higher for NP/S than NP/Z, but the human garden path effect patterned in the opposite direction (lower for NP/S), there is no single conversion factor that could bring the predictions of surprisal into alignment with the empirical results.

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 garden path effects and was unable to predict the relative difficulty of each construction compared with the others. Independently from the results of our computational simulations, a close inspection of the human reading times of words within each individual construction points to qualitative differences in the behavioral responses to the three constructions, again calling into question a uniform predictability-based account.

At a minimum, our results indicate that the relationship between surprisal and reading times is not linear in conditions such as the main verb / reduced relative ambiguity. 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, human sentence processing likely involves a syntactic repair mechanism (e.g., Sturt, 1997; Sturt et al., 1999) or a reprocessing mechanism (e.g., Grodner et al., 2003) that are invoked in challenging syntactic disambiguation contexts.

Acknowledgments

We are grateful to Brian Dillon, Roger Levy, Becky Marvin and Grusha Prasad, as well as audiences at the 2018 Annual Meeting of the Cognitive Science Society and AMLaP 2020, for engaging and helpful discussion regarding many aspects of this project, and to Dan Grodner and Nathaniel Smith for sharing their experimental materials. This work was supported in part by National Science Foundation grant BCS-2020945.

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 Fodor, J. and Ferreira, F., editors, *Reanalysis in sentence processing*, pages 1–46. Kluwer, Dordrecht.

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

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

Clifton Jr, C., Staub, A., and Rayner, K. (2007). Eye movements in reading words and sentences. In van Gompel, R. P. G., Fischer, M. H., Murray, W. S., and Hill, R. L., editors, *Eye Movements: A Window on Mind and Brain*, pages 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(3):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(2):193–210.

Demberg, V., Keller, F., and Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066.

Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, A. N. (2016). Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209. Association for Computational Linguistics.

Ehrlich, S. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

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

Enguehard, E., Goldberg, Y., and Linzen, T. (2017). Exploring the syntactic abilities of RNNs with multi-task learning. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 3–14.

Enochson, K. and Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PloS one*, 10(3):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(9):1362–1376.

Fine, A. B., Jaeger, T. F., Farmer, T. A., and Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLOS ONE*, 8(10):e77661.

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

Frank, S. and Hoeks, J. (2019). The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Socity (CogSci)*, pages 337–343. Cognitive Science Society.

Frank, S. L. (2010). Uncertainty reduction as a measure of cognitive processing effort. In Hale, J. T., editor, *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 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*. PhD 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(2):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(3):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)*, pages 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(1):58–93.

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

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Miyashita, Y., Marantz, A., and O'Neil, W., editors, *Image, language, brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT Press, Cambridge, MA.

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(20):8051–8056.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool, San Rafael, California.

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 the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

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

Grodner, D. J., Gibson, E., Argaman, V., and Babyonyshev, M. (2003). Against repair-based reanalysis in sentence comprehension. *Journal of Psycholinguistic Research*, 32(2):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)*, pages 1195–1205, New Orleans, Louisiana. 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*, pages 1–8, Pittsburgh, PA. Association for Computational Linguistics.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):609–642.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

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(2):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(2):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)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):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*, pages 234–243, Stroudsburg, PA. Association for Computational Linguistics.

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In van Gompel, R. P. G., editor, *Sentence Processing*, pages 78–114. Psychology Press.

Levy, R. P. and Keller, F. (2013). Expectation and locality effects in german verb-final structures. *Journal of Memory and Language*, 68(2):199–222.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):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(6):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(1):56–98.

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

Merkx, D. and Frank, S. L. (2020). Comparing Transformers and RNNs on predicting human sentence processing data. *arXiv preprint arXiv:2005.09471*.

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 Interna-*

tional Speech Communication Association (INTERSPEECH 2010), pages 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*, pages 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 Kieras, D. E. and Just, M. A., editors, *New Methods in Reading Comprehension Research*, pages 69–89. Erlbaum, Hillsdale, NJ.

Narayanan, S. and Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proceedings of the Twelfth Annual Meeting of the Cognitive Science Society*, pages 752–757.

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(4):786–803.

Padó, U., Crocker, M. W., and Keller, F. (2009). A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):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*.

Prasad, G. and Linzen, T. (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(3):539–576.

Rayner, K. and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3(4):504–509.

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. In

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 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*, pages 5387–5403, Online. Association for Computational Linguistics.

Shain, C. and Schuler, W. (2018). Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 2679–2689. Association for Computational Linguistics.

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

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201.

Sturt, P. (1997). *Syntactic reanalysis in human language processing*. PhD 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(5):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*, 30:415–433.

Van Petten, C. and Luka, B. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2):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)*, pages 2603–2608. Cognitive Science Society.

van Schijndel, M. and Linzen, T. (2019). Can entropy explain successor surprisal effects in reading? In Jarosz, G. and Pater, J., editors, *Proceedings of the 2nd Annual Meeting of the Society for Computation in Linguistics (SCiL)*, pages 1–7. Society for Computation in Linguistics, New York, NY.

van Schijndel, M., Schuler, W., and Culicover, P. W. (2014). Frequency effects in the processing of unbounded dependencies. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Socity (CogSci)*, pages 1658–1663. 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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wilcox, E., Levy, R., Morita, T., and Futrell, R. (2018). What do RNN language models learn about filler-gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., and Levy, R. (2019). Structural supervision improves learning of non-local grammatical dependencies. 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)*, pages 3302–3312, Minneapolis, Minnesota. Association for Computational Linguistics.