Modeling garden path effects without explicit hierarchical syntax

Marten van Schijndel (vansky@jhu.edu)
Department of Cognitive Science, Johns Hopkins University

Tal Linzen (tal.linzen@jhu.edu)
Department of Cognitive Science, Johns Hopkins University

Abstract

The disambiguation of syntactically ambiguous sentences can lead to reading difficulty, often referred to as a garden path effect. The surprisal hypothesis suggests that this difficulty can be accounted for using word predictability. We tested this hypothesis using predictability estimates derived from two families of language models: grammar-based models, which explicitly encode the syntax of the language; and recurrent neural network (RNN) models, which do not. Both classes of models correctly predicted increased difficulty in ambiguous sentences compared to controls, suggesting that the syntactic representations induced by RNNs are sufficient for this purpose. At the same time, surprisal estimates derived from all models systematically underestimated the magnitude of the effect, and failed to predict the difference between easier (NP/S) and harder (NP/Z) ambiguities. This suggests that it may not be possible to reduce garden path effects to predictability.

Keywords: self-paced reading; garden path; neural networks

Language is rife with temporary syntactic ambiguities. Most of these ambiguities go unnoticed during reading. In some cases, however, resolving the ambiguity can lead to substantial processing difficulty, as in the following example:

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

When the noun phrase *the instructor* is first read, it can be interpreted either as the object of *phoned* or as the subject of an upcoming clause. The disambiguating words *was very upset*, which rule out the direct object analysis, are typically read more slowly than they would be in an unambiguous sentence.

Temporary ambiguities as in (1) are said to lead the reader "down the garden path". Such garden path sentences have motivated a number of special-purpose reanalysis mechanisms that come into play in syntactically challenging circumstances (Fodor & Ferreira, 1998). A radically different proposal suggests that the words *was very upset* in (1) are read more slowly simply because they are unpredictable (Hale, 2001; Levy, 2008); specifically, a word is read more slowly the higher its *surprisal*, defined as follows:

$$S(w_i) = -\log_2 P(w_i \mid w_{1\ldots i-1}) \quad (1)$$

The surprisal hypothesis has at least two appealing properties. First, word predictability affects reading times even in the absence of syntactic ambiguity, and is therefore preferable on parsimony grounds to special reanalysis mechanisms. Second, surprisal estimates can be derived from any probability distribution over sequences of words (a *language model*); this makes it possible to use the variety of language models (LMs) implemented in the computational linguistics literature to make quantitative reading time predictions.

In a proof-of-concept demonstration using a fragment of English, Hale (2001) showed that surprisal from a grammar-based LM can qualitatively account for a particular case of syntactic disambiguation difficulty.[1] Later, surprisal from a broad-coverage grammar-based LM was shown to correlate with reading times (Demberg & Keller, 2008). To our knowledge, however, the hypothesis that disambiguation difficulty in garden path sentences can be reduced to surprisal has not been empirically tested with a broad-coverage LM.

We test this hypothesis using two types of LMs. First, grammar-based LMs, which consist of explicit syntactic representations and need to be trained on a large number of parsed sentences; and second, LMs based on recurrent neural networks (RNNs), which are not designed with symbolic internal representations of sentence structure and do not require syntactically annotated training data. While it is possible for hierarchical syntax to emerge in RNN LMs without explicit hierarchical syntax (Elman, 1991), and recent work has shown that such models are in fact fairly syntactically sophisticated (Linzen, Dupoux, & Goldberg, 2016; Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018), it is an empirical question whether the syntactic representations induced by an RNN are sufficient to produce predictability estimates that predict garden path effects in reading.

In the rest of this paper, we derive surprisal predictions for garden path sentences using broad-coverage grammar-based and RNN-based LMs, and examine whether explicitly modeling the grammar of the language provides an advantage in deriving the qualitative finding of increased processing difficulty in ambiguous sentences. Going beyond this qualitative question, we examine whether surprisal can predict the *magnitude* of disambiguation difficulty across two different types of ambiguity, NP/S and NP/Z.

Materials

We focus on two classic temporary ambiguities. The first type is the NP/S ambiguity, illustrated in (2a):

(2) 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.

[1] For work deriving surprisal estimates for temporarily ambiguous sentences from larger scale grammar-based LMs, see Levy (2013) and Linzen and Jaeger (2016).

This ambiguity is referred to as NP/S because *the contract* can initially serve either as a noun phrase (NP) complement to *understood* or as the subject of a sentential (S) complement. An unambiguous version of this sentence can be created by adding the overt complementizer *that*, as in (2b). Empirically, *would be changed* is read faster in (2b) in (2a).

The second ambiguity we investigate is the NP/Z ambiguity discussed in the introduction and repeated here as (3a):

(3) 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 (3a) are referred to as NP/Z sentences because the verb *phoned* is initially either transitive, with the noun phrase (NP) complement *the instructor*, or intransitive, with a "zero" (Z) complement. An unambiguous version of this sentence can be created by inserting a comma after the initial verb (3b); *had been drinking* is read faster in (3b) than in the ambiguous (3a). This ambiguity is often perceived to be harder to resolve than NP/S.

Modeling approach

We derived surprisal estimates from six LMs (described below) and evaluated them against the reading times (RTs) reported by Grodner, Gibson, Argaman, and Babyonyshev (2003). Grodner et al. collected word-by-word self-paced reading (SPR) data from 53 college-aged participants who read 20 ambiguous or unambiguous NP/S sentences, and 20 ambiguous or unambiguous NP/Z sentences.[2] They measured the total garden path effect, defined as the difference in RTs in the critical region between the ambiguous and unambiguous sentences; the critical region is *would be changed* in the NP/S case (2) and *was very upset* in the NP/Z case (3). We do not have access to item-by-item RTs since Grodner et al. only reported condition averages and confidence intervals; our modeling target is therefore the mean RT.

We averaged word surprisal over the critical region. If syntactic disambiguation difficulty is due entirely to the unpredictability of the disambiguating words, we expect one bit[3] of surprisal to have the same effect on RTs regardless of whether the word in question is in the disambiguating region or not. We can therefore use words from outside the disambiguating region to estimate the slowdown in milliseconds that each bit of surprisal causes, and convert our surprisal estimates to RT estimates using that multiplier. Since we did not have word-by-word RTs from Grodner et al. (2003), we estimate this multipler from Figure B2b in Smith and Levy (2013), which suggests that every bit of surprisal leads to a slowdown of approximately 3.75 milliseconds.

[2]Grodner et al. included both modified and unmodified variants of each sentence to test how the length of the ambiguous region affected the magnitude of the garden path effect; we focus here on the unmodified sentences, demonstrated in (2) and (3).

[3]Surprisal is typically measured in bits, reflecting the roots of this concept in information theory.

Grammar-based models

The grammar-based models all used lexicalized probabilistic context-free grammars (PCFGs).

Top-down parser: The Roark (2001) parser is trained on sections 02 to 21 of the Wall Street Journal (WSJ) portion of the Penn Treebank. Its surprisal estimates are frequently used in psycholinguistic studies and correlate well with RTs (Demberg & Keller, 2008; Roark, Bachrach, Cardenas, & Pallier, 2009).

Left corner parser: The van Schijndel, Exley, and Schuler (2013) left-corner parser is also trained on sections 02 to 21 of the WSJ corpus. This parser makes use of fine-grained grammatical distinctions induced by a split-merge procedure, which clusters syntactic categories based on the contexts in which they occur. We applied five iterations of split-merge to the grammar.

Left corner (categorial grammar): The van Schijndel et al. (2013) parser can be made even more context-sensitive by first reannotating the WSJ corpus with the Nguyen, van Schijndel, and Schuler (2012) generalized categorial grammar (GCG) annotation. This annotation produces a high degree of context-sensitivity reflecting deep syntactic dependencies similar to head-driven phrase structure grammar (HPSG) (Pollard & Sag, 1994). Three iterations of split-merge were applied to the annotated grammar.

Neural network models

The neural network models were all RNN models with long-short term memory (LSTM) units (Hochreiter & Schmidhuber, 1997) trained using PyTorch.[4] Unlike the grammar-based models, the RNNs were all trained exclusively on text without syntactic annotations.

Wall Street Journal: In order to directly compare the performance of RNNs to the PCFGs described above, one RNN was trained on Sections 02 to 21 of the Wall Street Journal corpus, just like the PCFGs.[5]

Wikipedia (2M words): For a larger training corpus, we trained an RNN on the Wikitext-2 corpus (Merity, Xiong, Bradbury, & Socher, 2016), using identical hyperparameters as were used with the Wall Street Journal RNN. Wikitext-2 contains around two million words of Wikipedia articles taken from the set of Good and Featured articles on Wikipedia.

Wikipedia (90M words): For an even larger training corpus, we used a model trained by Gulordava et al. (2018) on 90 million words extracted from English Wikipedia.[6]

[4]LSTM LM code which estimates surprisal and other incremental complexity measures is available at: `https://github.com/vansky/neural-complexity.git`

[5]We used the following hyperparameters: two LSTM layers with 1500 hidden units each, a 1500-dimensional input word embedding layer tied during training to the output weights, a dropout rate of 0.65, and trained for 40 epochs.

[6]This model had two LSTM layers with 650 hidden units each, 650-dimensional word embeddings, a dropout rate of 0.2 and batch size 128, and was trained for 40 epochs (with early stopping).

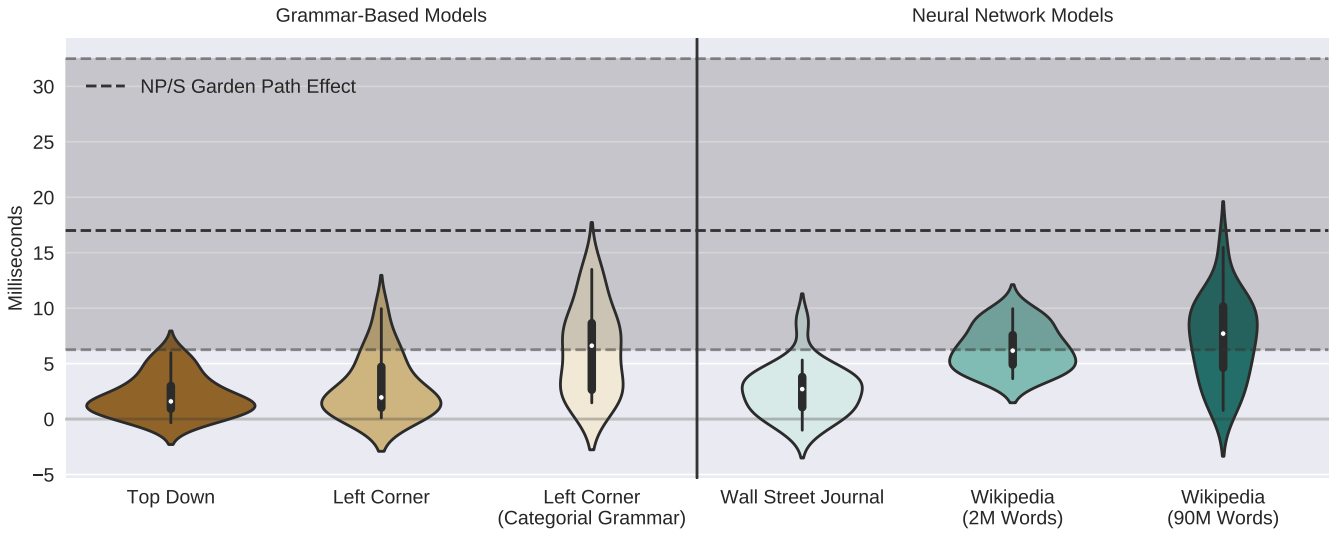

Figure 1: Reading time differences predicted by different LMs in the disambiguating region of ambiguous NP/S sentence compared to matched unambiguous controls (example (2) in the text). The violin plots show the distribution of predictions across items; the white dot indicates the median predicted difference. The empirical mean reading time difference reported by Grodner et al. (2003) is plotted with a dashed line;[7] the errors bars reported in that paper are represented using shading.

Results

Grodner et al. (2003) reported longer RTs in the critical region after an ambiguous prefix compared to an unambiguous one, both in NP/S sentences ((2a) compared to (2b) above) and in NP/Z sentences ((3a) compared to (3b)). All of our LMs correctly predicted that the critical region should be read more slowly in ambiguous than unambiguous sentences. One sample t-tests showed that the by-item mean surprisal difference in the critical region was significantly higher than 0 for all of the LMs for both NP/S and NP/Z.[8]

The magnitude of the predicted differences in NP/S sentences is shown in Figure 1. The median predictions of all models were significantly lower than the empirical mean RT difference, though for some models they were on the lower end of the confidence interval reported by Grodner et al. (2003). Among grammar-based LMs, this was only the case for the GCG LM; among RNN models, the model trained on 90M words predicted the largest difference, but the model trained on 2M words made comparable predictions.

There was a much greater discrepancy between the empirical and predicted RTs in NP/Z sentences (Figure 2). The empirical RT difference is much larger in NP/Z than in NP/S sentences (Sturt, Pickering, & Crocker, 1999; Grodner et al., 2003). The top-down and left-corner (non-GCG) LMs predicted a significantly larger NP/Z garden path effect compared with the NP/S effect, but this difference seems mainly due to the very low NP/S predictions of those models. The remaining models (with the exception of the 2M-word RNN) also predicted a numerically larger NP/Z effect which is not significant after correcting for multiple comparisons. Crucially, there was still a massive gap between the predictions of all of our LMs and even the lower end of the confidence interval for the empirical RTs; it appears unlikely that improvement in LM accuracy could substantially bridge this gap.

For both the NP/S and NP/Z cases, the overall distribution of predictions of the best models in each class (the GCG LM and the 90M-word RNN) did not differ in a meaningful way ($p = 0.98$). This suggests that RNNs can acquire sufficient syntactic knowledge to make predictions for syntactically ambiguous sentences that are similar to the predictions made by models with explicitly hierarchical structure.

Predictions for Sturt et al. (1999)

We next tested the models on the NP/S and NP/Z stimuli constructed by Sturt et al. (1999), which are similar in structure to those used by Grodner et al. (2003). The RTs predicted by our LMs were similar as well: the predicted NP/S effect size was 2-8 ms (compared to 2-7 ms for the Grodner materials) and the predicted NP/Z effect size was 5-10 ms (compared to 4-10 ms). Yet the discrepancy between predicted and empirical RTs was even larger in this case, because the empirical RT differences reported by Sturt are much larger than the effects reported by Grodner. While the NP/S garden path effect in Grodner is 17 ms, in Sturt it is 50 ms. Similarly, the NP/Z garden path in Grodner is 64 ms, while in Sturt it is a massive 152 ms (per word!).[9]

[7] Grodner et al. (2003) state that their mean NP/S effect size was 17 ms, but their plot with error bars shows a 20 ms mean NP/S effect. We use their plotted error bars but show their reported mean effect.

[8] We conducted 55 significance tests in this paper, so our corrected alpha value for significance was 0.0009. All effects we report as significant had p-values ≤ 0.0005.

[9] To make the analysis analogous to Grodner's word RTs, we divided Sturt's region RTs and our summed surprisals by the average

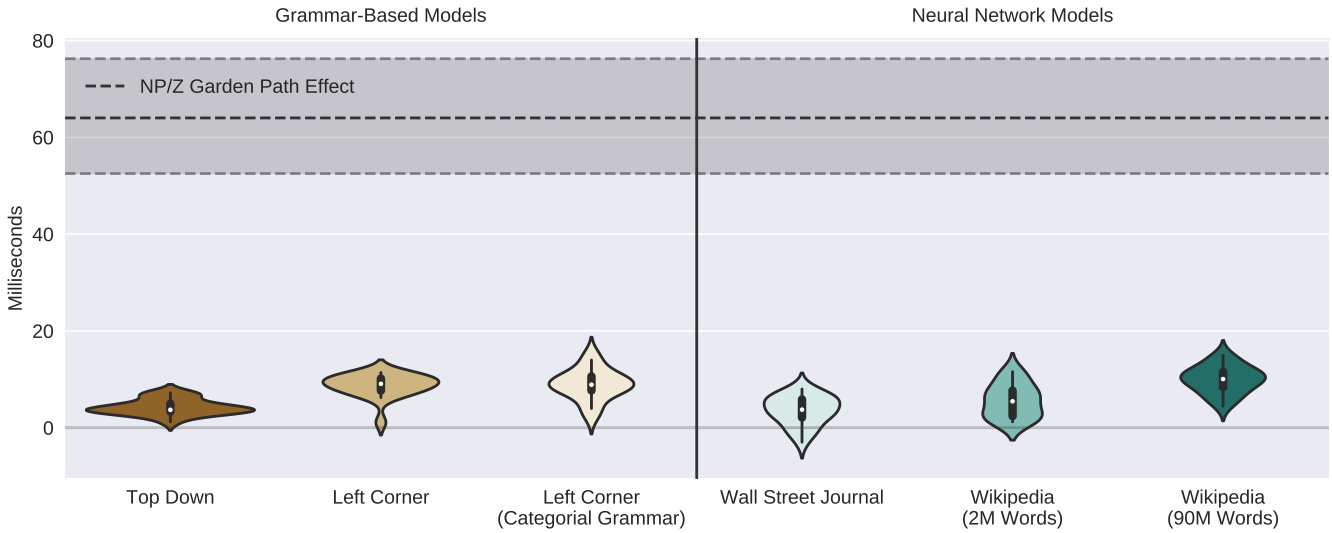

Figure 2: Reading time differences predicted by different LMs in the disambiguating region of ambiguous NP/Z sentence compared to matched unambiguous controls (example (3) in the text). The empirical mean reading time difference reported by Grodner et al. (2003) is plotted with a dashed line; the errors bars reported in that paper are represented using shading.

Our models made similar predictions across these two studies, suggesting that the empirical RT differences are unlikely to be due to differences in the lexical properties of the materials. We hypothesize that the source of the dramatic difference in RTs between the two studies is in the way the studies presented their stimuli. Whereas Grodner et al. used word-by-word self-paced reading, Sturt et al. used region-by-region self-paced reading. For example, an NP/Z sentence would be revealed region-by-region as follows:

(4) Before the woman / visited the famous doctor / **had been drinking** / quite a lot.

The region-by-region display likely encouraged subjects to chunk the sentence by regions, causing them to strongly commit to the ultimately incorrect parse in which the ambiguous noun phrase *the famous doctor* is the direct object of the preceding verb *visited*. The word-by-word display used by Grodner et al. appears to us to provide a more ecologically valid estimate of the RT cost of syntactic disambiguation.

Language model quality and fit to reading times

Goodkind and Bicknell (2018) found a linear relationship between the accuracy of a LM, as measured by its perplexity, and the extent to which surprisal estimates from the LM are predictive of RTs in naturally occurring sentences. In this section we explore whether this general finding holds for the disambiguating words in NP/S and NP/Z sentences.

We are unable to compare perplexity across the models described earlier since they had different vocabularies (due to the difference in corpus size). We therefore trained three new RNN LMs on the 2M words Wikipedia corpus, using three hyperparameter combinations suggested by the PyTorch developers (see Table 1). All training details were as before.

As expected, larger models obtained better (lower) perplexity. We derived surprisal estimates from all three models for the Grodner et al. NP/S materials. Despite the large change in linguistic accuracy (a 24% reduction in perplexity) across the models, there was not a correspondingly large change in the RT predictions (Figure 3). In fact, the mean NP/S garden path effect estimate of the most accurate model was numerically further from the human garden path effect than the least accurate model, though RT predictions are very similar across the models. While we acknowledge that even a 24% increase over a 6 ms difference is likely to be difficult to detect with a small number of items, we provisionally conclude that there is no evidence that the linear relationship that Goodkind and Bicknell (2018) found between linguistic accuracy and fit to psycholinguistic reading times extends to syntactically complex sentences.

Reading time predictions across the sentence

We have so far focused exclusively on the disambiguating region. We now briefly explore whether surprisal from our LMs

Model	Layers	Units	Dropout	Perplexity
Small	1	200	0.2	666.17
Medium	2	650	0.5	570.66
Large	2	1500	0.65	508.44

Table 1: RNN LMs evaluated in Figure 3. The Units column indicates the number of units in each layer. Perplexity was calculated on the Grodner et al. (2003) sentences.

number of words in each region (2.96 for the critical region).

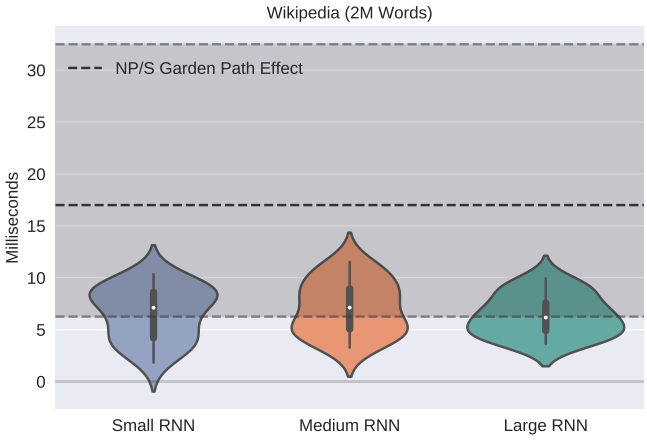

Figure 3: Effect of RNN hyperparameters (see Table 1) on reading time predictions for NP/S sentences.

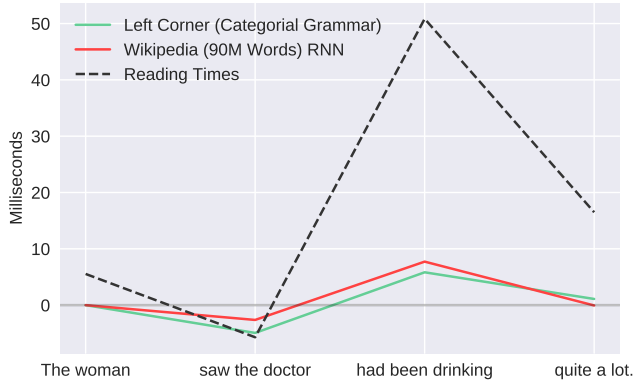

Figure 4: NP/S LM reading time predictions (surprisal converted to milliseconds); self-paced RT differences reported by Sturt et al. (1999) are plotted with a dashed line.

captures the reading patterns in the other regions of NP/S and NP/Z sentences. Grodner et al. (2003) did not report RTs for words after the disambiguating region, so for this analysis, we used the region-level self-paced RTs reported by Sturt et al. (1999), despite the caveats mentioned above. We focused on NP/S sentences because the immense NP/Z garden path effect reported by Sturt et al. (1999) distorted the scale of the graph, making interpretation difficult. We subtracted unambiguous NP/S RTs from ambiguous NP/S RTs and analyzed the region level differences between those conditions (see Figure 4).

In the first region, the two conditions are identical, so the models did not predict any difference between the conditions; likewise, Sturt et al. reported no significant difference. In the second region, reading was significantly slower in the unambiguous condition; most of our models were able to capture this pattern.[10] Sturt et al. speculated that the slower reading in this region may be driven by the additional *that* in the unambiguous condition, which the models capture by predicting some surprisal for the additional word. As mentioned earlier, the garden path effect observed by Sturt et al. is much greater than that reported by Grodner et al. (likely due to the region-level presentation), but all the models do correctly predicted a significant effect in this region. In the final region, the conditions are again identical, but the humans exhibit much slower reading in the ambiguous condition. Only the categorial grammar LM predicted significantly slower NP/S reading in the region, but all models except the 90M-word RNN predicted significantly slower NP/Z reading in the final region.

[10]The lone exception is the WSJ LSTM which predicted a non-significant mean effect of -1 ms in the second region.

Discussion

Classic explanations of disambiguation difficulty in temporarily ambiguous sentences have invoked special syntactic repair and reanalysis mechanisms. Surprisal theory raises the possibility that the elevated reading times at the disambiguating words are due to the low conditional probability (high surprisal) of those words in context. We have tested this hypothesis for two types of temporary syntactic ambiguities, NP/S and NP/Z, using two computational frameworks for estimating the conditional probability of words: grammar-based language models (LMs), estimated from collections of parse trees, and recurrent neural network (RNN) LMs, which are not explicitly constructed to represent syntax.

The best models in both classes correctly predicted that reading should be slower in the critical disambiguating region of ambiguous sentences compared to matched unambiguous sentences. Crucially, however, the predicted difference between the ambiguous and unambiguous sentences was on the same order of magnitude for the two types of ambiguities (NP/S and NP/Z). This conflicts with the empirical reading time effects, which are more pronounced in NP/Z than NP/S.

In addition to examining the qualitative pattern of predicted surprisal, we derived quantitative reading time predictions. We reasoned that if syntactic disambiguation difficulty is due to word predictability, we can estimate the reading slowdown caused by each bit of surprisal from reading times measured on sentences outside of the experimental materials. Since we did not have the full reading time data from the experiments we modeled (Grodner et al., 2003; Sturt et al., 1999), we derived our estimate from a reading-time corpus study (Smith & Levy, 2013). Based on this estimate, we found that even our best models underestimated the NP/S effect by a factor of more than two (a predicted 6-7 ms slowdown compared to the empirical 17 ms, although the estimate of 6-7 ms is within the errors bars for the effects). Surprisal from the LMs underestimated the NP/Z effect by a much larger factor.

It is certainly possible that surprisal had a larger effect per bit in the experiments we modeled than in the Smith and Levy (2013) reading time corpus; for example, the higher proportion of syntactically complex sentences could have caused participants to read more slowly overall. If the true slowdown per bit of surprisal was indeed higher than our estimate, the predictions made by our best LMs for the NP/S case could get

closer to the empirical effect size, and the NP/S disambiguation effect could indeed reduce to word predictability.

This would not be sufficient for the NP/Z ambiguity, however; for surprisal to account for both this case and the NP/S case, a LM would need to make dramatically different predictions from the ones that our best LMs made. Given the high linguistic accuracy of contemporary RNN LMs, it is more likely that surprisal alone simply cannot account for NP/Z disambiguation difficulty; an additional mechanism is required, perhaps along the lines of the special syntactic reanalysis mechanisms proposed in Fodor and Ferreira (1998) (see also Sturt et al., 1999). It is of course possible that any such additional syntactic mechanisms are at play in NP/S disambiguation as well, but have a less dramatic effect.

Frank and Christiansen (2018) hypothesized that explicit hierarchical syntax and the associated parsing algorithms may not be necessary for modeling sentence processing. The present findings provide partial support for their hypothesis, in that reading times in syntactically ambiguous sentences were predicted comparably by grammar-based and neural LMs.[11] Importantly, our results should not be taken to suggest that the RNN LMs do not *induce* hierarchical syntactic representations—given that the phenomena they need to model are hierarchical, and given that their linguistic accuracy in practice is high, it is very likely that they do acquire hierarchical representations, although those representations may be imperfect (Linzen et al., 2016). Our results only indicate that such representations do not need to be built into the architecture of the neural network, and do not need to be presented to the network during training, to obtain comparable predictions to grammar-based models. It remains to be seen if the additional syntactic repair mechanisms that appear to be necessary to account for NP/Z disambiguation difficulty can be expressed in a neural network, or whether they require an explicit syntactic representation (Sturt et al., 1999).

Acknowledgments

Thanks to Grusha Prasad and Becky Marvin for engaging and helpful discussion regarding many aspects of this project, to Dan Grodner for sharing his experimental materials, and to Kristina Gulordava for sharing her LSTM LM with us.

References

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

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

Fodor, J., & Ferreira, F. (Eds.). (1998). *Reanalysis in sentence processing*. Dordrecht: Kluwer.

[11] A direct comparison between grammar-based models and RNNs is difficult. RNN LMs may be better able to capture fine-grained selectional preferences, especially when trained on 90 million words, which are likely to be central to modeling garden path effects (Levy, 2013).

Frank, S. L., & Christiansen, M. H. (2018). Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience*. doi: 10.1080/23273798.2018.1424347

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

Grodner, D. J., Gibson, E., Argaman, V., & 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., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL.*

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

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 78–114). Hove: Psychology Press.

Linzen, T., Dupoux, E., & 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., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 1–30.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). *Wikitext-2* (Tech. Rep.). Salesforce.

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

Pollard, C., & Sag, I. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, *27*(2), 249–276.

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

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

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

van Schijndel, M., Exley, A., & Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, *5*(3), 522–540.