A Neural Model of Adaptation in Reading

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

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

Abstract

It has been argued that humans rapidly adapt their lexical and syntactic expectations to match the statistics of the current linguistic context. We provide further support to this claim by showing that the addition of a simple adaptation mechanism to a neural language model improves our predictions of human reading times compared to a non-adaptive model. We analyze the performance of the model on controlled materials from psycholinguistic experiments and show that it adapts not only to lexical items but also to abstract syntactic structures.

1 Introduction

Reading involves the integration of noisy perceptual evidence with probabilistic expectations about the likely contents of the text. Words that are consistent with these expectations are identified more quickly (Ehrlich and Rayner, 1981; Smith and Levy, 2013). For the reader's expectations to be maximally effective, they should not only reflect the reader's past experience with the language (Hale, 2001; MacDonald and Christiansen, 2002), but should also be *adapted* to the current context. Optimal adaptation would reflect properties of the text being read, such as genre, topic and writer identity, as well as the general tendency for recently used words and syntactic structures to be reused with higher probability (Bock, 1986; Church, 2000; Dubey et al., 2006).

Several studies have suggested that readers do in fact adapt their lexical and syntactic predictions to the current context (Otten and Van Berkum, 2008; Fine et al., 2013; Fine and Jaeger, 2016).[1] For example, Fine and Jaeger investigated the processing of "garden path" sentences such as (1):

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

The word *warned* in (1) is initially ambiguous between a main verb interpretation (the soldiers were doing the warning) and a reduced relative clause interpretation (the soldiers were being warned). When the word *conducted* is reached, this ambiguity is resolved in favor of the reduced relative parse. Reduced relatives are infrequent constructions. This makes the disambiguating word *conducted* unexpected, causing it to be read more slowly than it would be in a context such as (2), in which the words *who were* indicate early on that only the relative clause parse is possible:

(2) The experienced soldiers who were warned about the dangers conducted the midnight raid.

Fine and Jaeger included a large proportion of reduced relatives in their experiment. As the experiment progressed, the cost of disambiguation in favor of the reduced relative interpretation decreased, suggesting that readers had come to expect a construction that is normally infrequent.

Human syntactic expectations have been successfully modeled with syntax-based language models (Hale, 2001; Levy, 2008; Roark et al., 2009). Recently, language models (LMs) based on recurrent neural networks (RNNs) have been shown to make adequate syntactic predictions (Linzen et al., 2016; Gulordava et al., 2018), and to make comparable reading time predictions to syntax-based LMs (van Schijndel and Linzen, 2018). In this paper, we propose a simple way to continuously adapt a neural LM, and test the method's psycholinguistic plausibility. We show that LM adaptation significantly improves our ability to predict human reading times using the LM. Follow-up experiments with controlled materials show that the LM adapts not only to specific

[1] Recently, Harrington Stack et al. (2018) questioned the robustness of the results of Fine et al. (2013).

vocabulary items but also to abstract syntactic constructions, as humans do.

2 Method

We use a simple method to adapt our LM: at the end of each new test sentence, we update the parameters of the LM based on its cross-entropy loss when predicting that sentence; the new weights are then used to predict the next test sentence.[2] Our baseline LM is a long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997) language model trained on 90 million words of English Wikipedia by Gulordava et al. (2018) (see Supplementary Materials for details). For adaptation, we keep the learning rate of 20 used by Gulordava et al. (the gradient is multiplied by this learning rate during weight updates). We examine the effect of this parameter in Section 5.2.

We tested the model on the Natural Stories Corpus (Futrell et al., 2018), which has 10 narratives with self-paced reading times from 181 native English speakers. There are two narrative genres in the corpus: fairy tales (seven texts) and documentary accounts (three texts).

3 Linguistic accuracy

We first measured how well the adaptive model predicted upcoming words. We report the model's perplexity, a quantity which is lower when the LM assigns higher probabilities to the words that in fact occurred. We adapted the model to the first k sentences of each text, then tested it on sentence $k+1$, for all k. Adaptation dramatically improved test perplexity compared to the non-adaptive version of the model (86.99 vs. 141.49).

We next adapted the model to each genre separately. If the model adapts to stylistic or syntactic patterns, we might expect adaptation to be more helpful in the fairy tale than the documentary genre: the Wikipedia corpus that the LM was originally trained on is likely to be more similar in style to the documentary genre. Consistent with this hypothesis, the documentary texts benefited less from adaptation (99.33 to 73.20) than the fairy tales (160.05 to 86.47), though the fact that both saw improvement from adaptation suggests that text-specific adaptation is beneficial even if the genre is similar to the training genre.

[2]Our code is publicly available at: https://github.com/vansky/adaptive-LM.git

	$\hat{\beta}$	$\hat{\sigma}$	t
WITHOUT ADAPTIVE SURPRISAL:			
Sentence position	0.55	0.53	1.03
Word length	7.29	1.00	7.26
Non-adaptive surprisal	6.64	0.68	9.79
WITH ADAPTIVE SURPRISAL:			
Sentence position	0.29	0.53	0.55
Word length	6.42	1.00	6.40
Non-adaptive surprisal	-0.89	0.68	-1.31
Adaptive surprisal	8.45	0.63	13.42

Table 1: Fixed effect regression coefficients from fitting self-paced reading times. The top model lacks fixed and random effects of adaptive surprisal. In general, a coefficient is significant when $|t| > 2$.

Each genre consists of multiple texts. Does adaptation to a particular text lead to catastrophic forgetting (McCloskey and Cohen, 1989), such that the LM overfits to the text and forgets its more general knowledge acquired from the Wikipedia training corpus? This was not the case; in fact, adapting to the entirety of each genre without reverting to the baseline model after each text led to a very slightly *better* perplexity (fairytales: 86.47, documentaries: 73.20) compared with a setting in which the LM was reverted after each text (fairytales: 86.61, documentaries: 73.63).

4 Modeling human expectations

We next tested whether our adaptive LM matches human expectations better than a non-adaptive model. Since each reader saw the texts in a different order, we adapted the LM to each text separately: after each story, we reverted to the initial Wikipedia-trained LM and restarted adaptation on the next text. If anything, this likely resulted in a conservative estimate of the benefit of adaptation compared to a model that adapts continuously across multiple stories from the same genre, as humans might do.[3]

We used surprisal as a linking function between the LM's predictions and human reading times

[3]We do not distinguish between *priming* and *adaptation* in this paper. While it may be tempting to think of the LSTM memory cell as a model of priming and of the weight updates as a model of adaptation, Bock and Griffin (2000) provide evidence that priming cannot simply be a function of residual activation and that priming can be driven by longer-term learning (see Tooley and Traxler (2010) for more discussion on priming vs. adaptation).

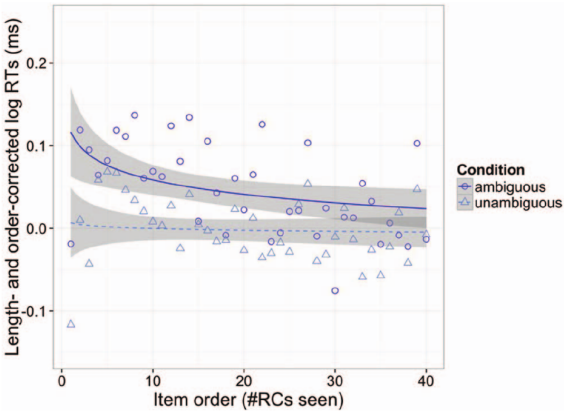

Figure 1: Mean length- and order-corrected reading times over the disambiguating region of the critical items in Fine and Jaeger (2016). Figure adopted from that paper.

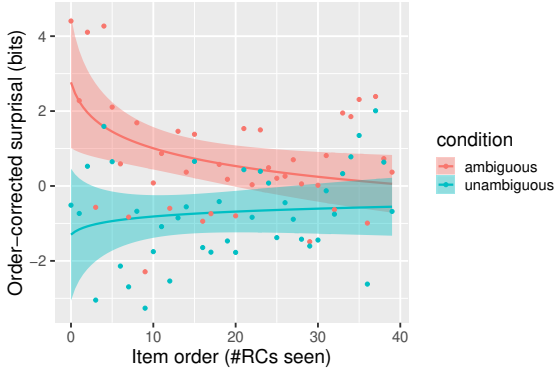

Figure 2: Mean order-corrected model surprisal over the disambiguating region of the critical items in Fine and Jaeger (2016).

(Hale, 2001; Smith and Levy, 2013). Surprisal quantifies how unpredictable each word (w_i) is given the preceding words:

$$\text{surprisal}(w_i) = -\log \text{P}(w_i \mid w_1 ... w_{i-1}) \quad (1)$$

We fit the self-paced reading times in the Natural Stories Corpus with linear mixed effects models (LMEMs), a generalization of linear regression (see Supplementary Materials for details).

In line with previous work, non-adaptive surprisal was a significant predictor of reading times ($p < 0.001$) when the model only included other baseline factors (Table 1, Top). Adaptive surprisal was a significant predictor of reading times ($p < 0.001$) over non-adaptive surprisal and all baseline factors (Table 1, Bottom). Crucially, non-adaptive surprisal was no longer a significant predictor of reading times once adaptive surprisal was included. This indicates that the predictions of the adaptive model subsume the predictions of the non-adaptive one.

5 Does the model adapt to syntax?

We have shown that LM adaptation improves our ability to model human expectations as reflected in a self-paced reading time corpus. How much of this improvement is due to adaptation of the model's syntactic representations (Bacchiani et al., 2006; Dubey et al., 2006) and how much is simply due to the model assigning a higher probability to words that have recently occurred (Kuhn and de Mori, 1990; Church, 2000)? We address this question using two syntactic phenomena: reduced relative clauses and the dative alternation.

5.1 Reduced relative clauses

We adapted the model independently to random orderings of the critical and filler stimuli used in Experiment 3 of Fine and Jaeger (2016);[4] this experiment (described in the Introduction) contained a much higher proportion of reduced relative clauses than their general distribution in English. We used surprisal as our proxy for reading times. Following Fine and Jaeger, we took the mean surprisal over three words in each ambiguous sentence: the disambiguating word and the following two words (e.g., *conducted the midnight* in example (1)). To estimate the magnitude of the syntactic disambiguation penalty while also controlling for lexical content, we subtracted this quantity from the mean surprisal over the exact same words in the paired unambiguous sentence (2). Linear regression showed that the disambiguation penalty decreased as the model was exposed to more critical items (item order coefficient: $\hat{\beta} = -0.0804$, $p < 0.001$), indicating that the LM was adapting to reduced relatives, a syntactic construction without any lexical content.

In order to compare our findings more directly with the results given by Fine and Jaeger (2016) (shown in Figure 1), we mimicked their method of plotting reading times. First, we fit a linear model of the mean surprisal of each disambiguating region with the number of trials the model had seen in the experiment thus far to account for a general trend of subjects speeding up over the course

[4] See details in the Supplementary Materials.

of the experiment. Then, we plotted the mean residual model surprisal that was left in the disambiguating region in both the ambiguous and unambiguous conditions as the experiment progressed. The shape of our model's adaptation to the reduced relative construction (upper curve in Figure 2) matched the human results reported by Fine and Jaeger. Like humans, the model showed an initially large adaptation effect, followed by more gradual adaptation thereafter. Both humans and our model continued to adapt over all the items rather than just at the beginning of the experiment. Also like humans, the model's response to unambiguous items did not change significantly over the course of the experiment ($p = 0.91$).

5.2 The dative alternation

Dative events can be expressed using two roughly equivalent English constructions:

(3) a. *Prepositional object (PO)*:
The boy threw the ball to the dog.
b. *Double object (DO)*:
The boy threw the dog the ball.

Work in psycholinguistics has shown that recent experience with one of these variants increases the probability of producing that variant (Bock, 1986; Kaschak et al., 2006) as well as the likelihood of predicting it in reading (Tooley and Bock, 2014). To test whether our adaptation method can reproduce this behavior, we generated 200 pairs of dative sentences similar to (3). We shuffled 100 DO sentences into 1000 filler sentences sampled from the Wikitext-2 training corpus (Merity et al., 2016) and adapted the model to these 1100 sentences. We then froze the weights of the adapted model and tested its predictions for two types of sentences: the PO counterparts of the DO sentences in the adaptation set, which shared the vocabulary of the adaptation set but differed in syntax; and 100 new DO sentences, which shared syntax but no content words with the adaptation set.[5]

An additional goal of this experiment was to examine the effect of learning rate on adaptation. During adaptation the model performs a single parameter update after each sentence and does not train until convergence with gradual reduction of the learning rate as would normally be the case during LM training. Consequently, the learning rate parameter crucially determines the amount of adaptation the model can undertake after each sentence. If the learning rate is very low, adaptation will not have any effect; if it is too high, either the model will overfit after each update and will not generalize well, or the model will forget its trained representation as it overshoots the targeted minima. The optimal rate may differ between lexical and syntactic adaptation. Our experiments thus far all used the same learning rate as our original model (20); here, we varied the learning rate on a logarithmic scale between 0.002 and 200.

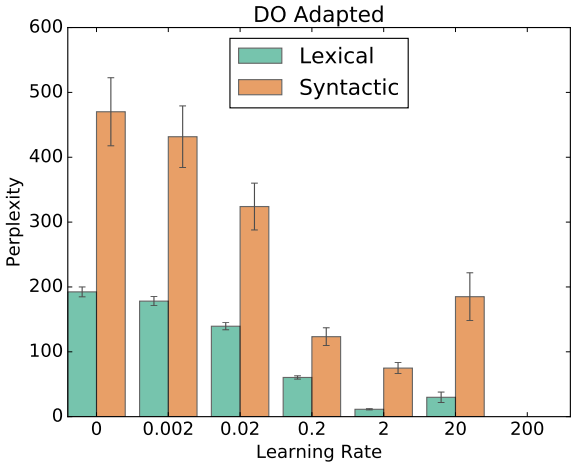

Figure 3: Learning rate influence over syntactic and lexical adaptation. A learning rate of 0 indicates the performance of the non-adaptive model; the learning rate of 200 resulted in perplexity in the billions.

The results of this experiment are shown in Figure 3. The model successfully adapted to the DO construction as well as to the vocabulary of the adaptation sentences. This was the case for all of the learning rates except for 200, which resulted in enormous perplexity on both sentence types. Both lexical and syntactic adaptation were most successful when the learning rate was around 2, with perplexity reductions of 94% for lexical adaptation and 84% for syntactic adaptation.

The better perplexity for lexical adaptation (testing on PO) compared with syntactic adaptation (testing on DO) arose from the larger number of PO sentences in the pre-training corpus. Syntactic adaption was penalized at higher learning rates more than lexical adaptation (compare learning rates of 2 and 20). This fragility of syntactic adaptation likely stems from the fact that the model can directly observe the relevant vocabu-

[5] For additional details as well as the reverse setting (adaptation to PO), see Supplementary Materials.

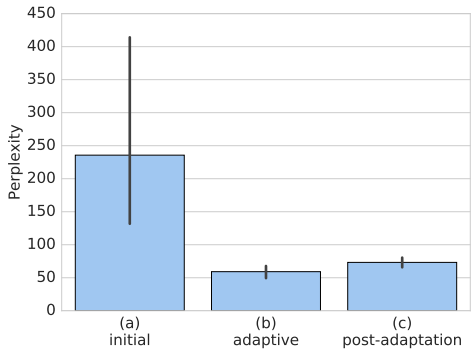

Figure 4: Perplexity on (a) the adaptation set of *G1* before adaptation, (b) the adaptation set of *G1* during the adaptation process, (c) the held-out set of *G1* after adapting to *G1* then adapting to *G2*.

lary, but syntax is latent and must be inferred from multiple similar sentences, a generalization which is impeded by overfitting at higher learning rates.

6 Testing for catastrophic forgetting

Our analysis of the Natural Stories corpus did not indicate that the model suffered from catastrophic forgetting. Yet the Natural Stories corpus contained only two genres; to address the issue of catastrophic forgetting more systematically, we used the premise sentences from the MultiNLI corpus (Williams et al., 2018) — a total of 2000 sentences for each of 10 genres.

For each genre pair G_1 and G_2, we first adapted the baseline model trained on Wikipedia to 1000 sentences of G_1 using a learning rate of 2 (shown to be optimal in Section 5.2). We then adapted the model to 1000 sentences of G_2. Finally, we froze the model's weights and tested its perplexity on the 1000 held out sentences from G_1.

The results averaged across all pairs of genres are plotted in Figure 4. Unsurprisingly, the model performed best on G_1 immediately after adapting to it (middle bar). Crucially, even after adapting to 1000 sentences of G_2 after its last exposure to G_1, it still modeled G_1 (right bar) much better than the non-adapted model (left bar). These results suggest that catastrophic forgetting is not a concern even with a relatively large amount of data.

7 Discussion

Adaptation greatly improved an RNN LM's word prediction accuracy, in line with other work on LM adaptation (Kneser and Steinbiss, 1993). We showed that the adapted model was psycholinguistically plausible, in two senses. First, it improved the correlation between surprisal derived from the model and human reading times, suggesting that the model generated more human-like expectations. Second, using materials that teased apart lexical content from syntax, we showed that the model adapted both its lexical and its syntactic predictions, in line with findings from human experiments. Finally, as in other neural-network based models in psychology (Chang et al., 2006), our gradient-based updates naturally incorporate the error-driven nature of syntactic adaptation; while we did not demonstrate this in the current paper, we hypothesize that our model will reproduce the finding that more surprising words lead to greater adaptation (Jaeger and Snider, 2013).

The simplicity of our adaptation method makes it attractive for use in modeling human expectations. Since adaptive surprisal is strictly superior to non-adaptive surprisal in modeling reading times, it would be a stronger baseline in analyses that aim to demonstrate the contribution of factors other than predictability.

We used a simple neural adaptation approach, where we performed continuous gradient updates based on the prediction error on the adaptation sentences (see also Krause et al., 2017). An alternative approach to neural LM adaptation uses recent RNN states in conjunction with the current state to make word predictions (Grave et al., 2017; Merity et al., 2017); a comparison of the two methods using our paradigms may provide insight into their relative strengths and weaknesses.

Finally, we reverted to the base model after the end of each text in our experiments, forgetting any text-specific adaptation. This mimics the effect of a participant leaving an experiment that had an unusual distribution of syntactic constructions and reverting to their standard expectations. In practice, however, humans are able to generalize from prior experience when they begin adapting to a new speaker or text if it is similar in some way to their previous experiences. For example, the model of Jaech and Ostendorf (2018) adapts to environmental factors, so it could potentially draw on independent experiences with female speakers and with lawyer speech in order to initialize a model of adaptation to a new female lawyer (see also Mikolov and Zweig, 2012; Kleinschmidt, 2018). The psycholinguistic plausibility of these models can be tested in future work.

References

Michiel Bacchiani, Michael Riley, Brian Roark, and Richard Sproat. 2006. MAP adaptation of stochastic grammars. *Computer Speech & Language*, 20(1):41–68.

Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Kathryn Bock and Zenzi M. Griffin. 2000. The persistence of structural priming: transient activation or implicit learning? *Journal of Experimental Psychology: General*, 129(2):177–192.

Franklin Chang, Gary S Dell, and Kathryn Bock. 2006. Becoming syntactic. *Psychological Review*, 113(2):234–272.

Kenneth W. Church. 2000. Empirical estimates of adaptation: the chance of two noriegas is closer to $p/2$ than p^2. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, pages 180–186. Association for Computational Linguistics.

Amit Dubey, Frank Keller, and Patrick Sturt. 2006. Integrating syntactic priming into an incremental probabilistic parser, with an application to psycholinguistic modeling. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics.

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

Alex B. Fine and T. Florian Jaeger. 2016. The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9):1362–1376.

Alex B. Fine, T. Florian Jaeger, Thomas A. Farmer, and Ting Qian. 2013. Rapid expectation adaptation during syntactic comprehension. *PloS ONE*, 8(10):1–18.

Richard Futrell, Edward Gibson, Hal Tily, Anastasia Vishnevetsky, Steve Piantadosi, and Evelina Fedorenko. 2018. The natural stories corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 76–82.

Edouard Grave, Armand Joulin, and Nicolas Usunier. 2017. Improving neural language models with a continuous cache. In *Proceedings of the International Conference on Learning Representations (ICLR 2017)*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 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. Association for Computational Linguistics.

John Hale. 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 Human Language Technologies*, pages 1–8, Pittsburgh, PA. Association for Computational Linguistics.

Caoimhe M. Harrington Stack, Ariel N. James, and Duane G. Watson. 2018. A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*.

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

Aaron Jaech and Mari Ostendorf. 2018. Personalized language model for query auto-completion. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 700–705. Association for Computational Linguistics.

T. Florian Jaeger and Neal E. Snider. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127:57–83.

Michael P. Kaschak, Renrick A. Loney, and Kristin L Borreggine. 2006. Recent experience affects the strength of structural priming. *Cognition*, 99(3):B73–B82.

Dave F. Kleinschmidt. 2018. Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*.

Reinhard Kneser and Volker Steinbiss. 1993. On the dynamic adaptation of stochastic language models. In *Proceedings of ICASSP-93*.

Ben Krause, Emmanuel Kahembwe, Iain Murray, and Steve Renals. 2017. Dynamic evaluation of neural sequence models. *arXiv preprint arXiv:1709.07432*.

Roland Kuhn and Renato de Mori. 1990. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583.

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

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Maryellen C. MacDonald and Morten H. Christiansen. 2002. Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, 109(1):35–54.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, *The Psychology of Learning and Motivation: Volume 24*, 92, pages 109–165. San Diego: Academic Press.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Wikitext-2.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of the International Conference on Learning Representations (ICLR 2017)*.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. *SLT*, 12:234–239.

Marte Otten and Jos J. A. Van Berkum. 2008. Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, 45(6):464–496.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 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. Association for Computational Linguistics.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Kristen M. Tooley and Kathryn Bock. 2014. On the parity of structural persistence in language production and comprehension. *Cognition*, 132(2):101–136.

Kristen M. Tooley and Matthew J. Traxler. 2010. Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*, 4(10):925–937.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In Tim Rogers, Marina Rau, Jerry Zhu, and Chuck Kalish, editors, *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 2600–2605. Cognitive Science Society, Austin, TX.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.