Using syntactic priming to investigate how recurrent neural networks represent syntax

Grusha Prasad, Marten van Schijndel and Tal Linzen *(Johns Hopkins University)*
grusha.prasad@jhu.edu

Introduction: Recent studies have demonstrated that recurrent neural networks (RNNs) trained to process sentences encode some syntactic information (Gulordava et al, 2018; Futrell et al, 2018). This raises the possibility that these networks can serve as a component in computational models of human sentence processing. In order for that to happen, we need to better understand what specific syntactic information these networks encode. We propose a paradigm based on syntactic priming that allows us to investigate an RNN's internal grammar and apply it to a specific RNN to understand how it represents relative clauses (RCs).

Branigan (1995) argued that syntactic priming — the phenomenon where a sentence becomes easier to process if it is preceded by a sentence with the same syntactic structure — can be used to infer human syntactic representations. van Schijndel & Linzen (2018) (vS&L) found a phenomenon analogous to priming in comprehension with RNNs. When a fully trained RNN was presented with a small number of sentences with a shared syntactic structure, and was allowed to update its parameters at the end of every sentence (i.e. *adapt* to the sentence), the surprisal (negative log probability) for novel sentences with that structure decreased. Therefore, we can use this adaptive mechanism to infer an RNN's syntactic representations. If the surprisal for an RNN on sentences with a syntactic structure K is lower when adapted to N sentences with structure K than when adapted to N fillers, then the RNN likely represents K.

As a proof of concept, we applied this method to investigate whether an RNN is sensitive to the structural similarity between reduced and unreduced RCs (RRC & URC) like in (1a) & (1b).

1(a) RRC: [The six volunteers taught the complicated procedure] learned it very well.

(b) URC: [The six volunteers who were taught the complicated procedure] learned it very well.

vS&L found that as an RNN adapted to RRCs, the surprisal at the disambiguating verb (underlined) for (1a) decreased more rapidly than the surprisal for (1b). We use this decrease in surprisal as our dependent measure in the following experiments.

Models: We created 12 clones of an RNN language model trained on 2 million words of English Wikipedia (wRNN). Each clone was adapted to either 12 RRCs (RRC-adapted), 12 URCs (URC-adapted) or 12 Fillers (Filler-adapted). All the RCs had ditransitive verbs and adjectives.

Experiment 1: To test whether this RNN (wRNN) was sensitive to the structural similarity between RRCs and URCs, we tested all the adapted clones of the wRNN on two sets of 12 sentences whose RRCs varied syntactically from the RRCs in the adaptation set.

2. *No adjectives*: [The girl bought the toy] played with it all day.

3. *No ditransitive verb*: [The lions attacked during the day] were unable to escape the hunters.

The surprisal at the disambiguating verb (underlined) for the RRC and URC adapted clones was lower than that for the Filler adapted clone (see Figure 1), suggesting that wRNN was sensitive to the structural similarity between RRCs and URCs.

Experiment 2: We wanted to test whether the results of Experiment 1 could have been driven by a simple heuristic the clones learned (e.g., sentences frequently have two past tense verbs). We tested all the adapted clones on 12 sentences like (4), that had two past tense verbs but no RC. If they just learned the heuristic, all the clones should have equivalent surprisal.

(4) The girl bought the toy and saw the movie.

The surprisal at *saw* and the period for the RRC and URC adapted clone was greater than that for the Filler adapted clone (Figure 2), suggesting that the wRNN did not just learn the heuristic.

Conclusion: We have proposed a paradigm inspired by syntactic priming that allows us to investigate an RNN's syntactic representations. Applying this paradigm, we found that the RNN we tested (wRNN) was sensitive to the structural similarity between URCs and RRCs.

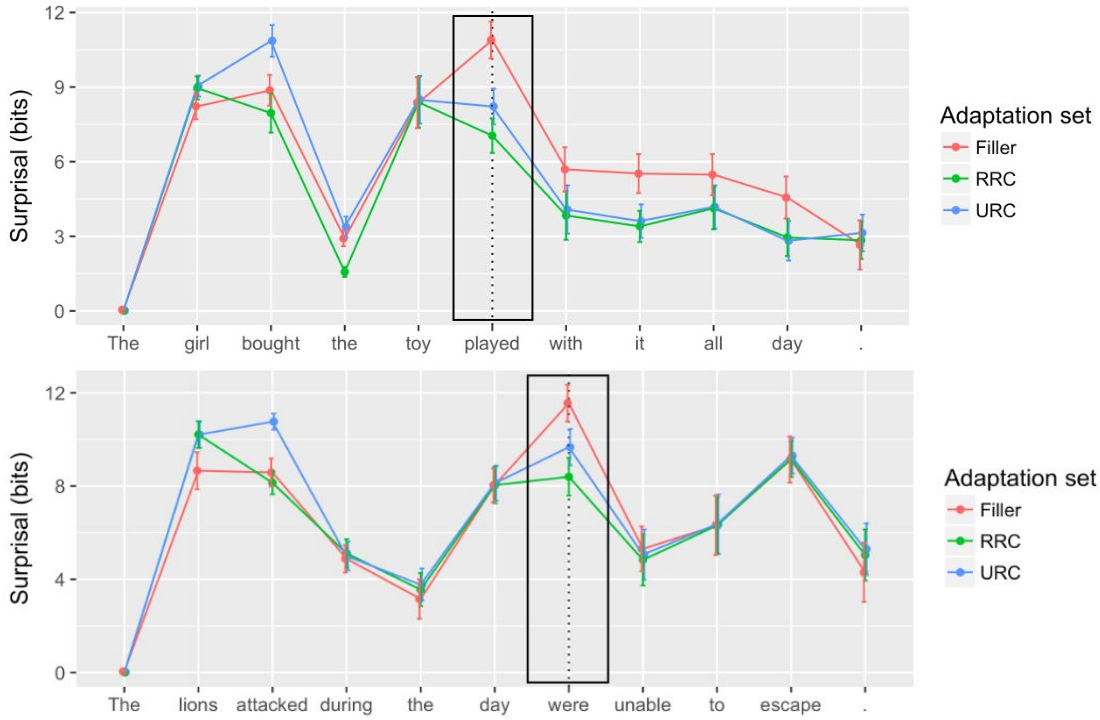

Figure 1: Surprisal across the sentence in Experiment 1 (example 2 top; example 3 bottom) for models adapted to different sets, averaged across 4 random orders. Paired t-tests revealed that at the disambiguating verb (highlighted) the surprisal for Filler > URC > RRC ($p << 0.001$). The surprisal at the 3rd word was greater for URC than for Filler or RRC. This is likely driven by the URC model strongly predicting a relativizer and is not relevant to our question. Error bars represent 95% CIs.

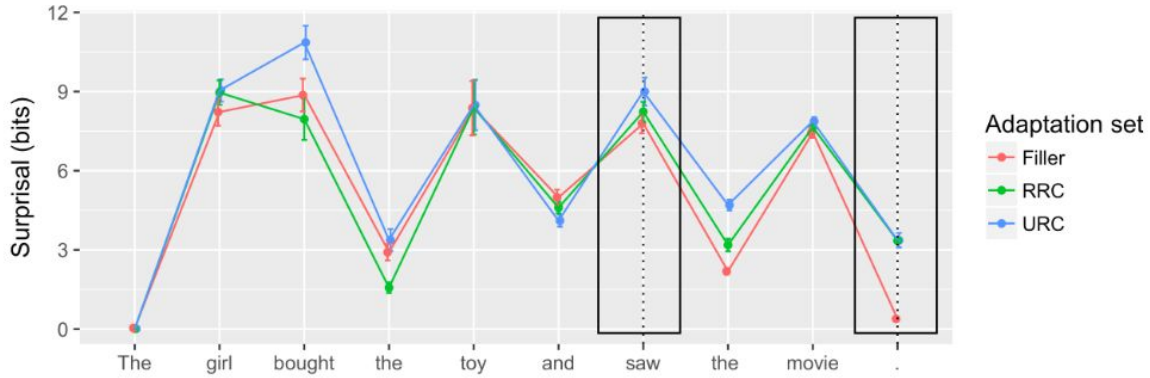

Figure 2: Surprisal across the sentence in Experiment 2 for models adapted to different sets, averaged across the same 4 random orders as in Experiment 1. Paired t-tests revealed that at *saw* surprisal for RRC > Filler ($p < 0.05$) and URC > RRC ($p << 0.001$) and at the period RRC = URC > Filler ($p << 0.001$)

References:

- Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. J., & Urbach, T. P. (1995). Syntactic priming: Investigating the mental representation of language. Journal of Psycholinguistic Research, 24(6), 489-506.
- Futrell, R., Wilcox, E., Morita, T., & Levy, R. (2018). RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. arXiv preprint arXiv:1809.01329.
- Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In Proceedings of NAACL-HLT, pages 1195–1205
- van Schijndel, M., & Linzen, T. (2018). A Neural Model of Adaptation in Reading. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)