Structure in Training Data Matters Less Than Expected: Surprisal From Some Scrambled Language Models Predicts Naturalistic Reading Times

Iza Škrjanec (skrjanec@lst.uni-saarland.de)[1]
[1]Saarland University

Language models (LM) have been proposed as computational models for estimating the processing effort during reading related to predictability, most notably via surprisal [3]. Neural LMs have been proven useful, providing surprisal estimates that predict reading times (RT) of naturalistic text [e.g., 8]. However, surprisal from neural LMs seems to systematically underpredict processing difficulty in complex constructions [e.g., 4]. In this work, we test whether structural properties of LM training data affect surprisal's ability to predict naturalistic RT. Our main question is: if a model trained on linguistically implausible data still predicts RT, what does this reveal about what LM surprisal is actually measuring?

Following [5, 9], we train English GPT2-small LMs from scratch under three conditions: a control condition trained on natural English, and two scrambled conditions: local bigram scramble, and global sentence scramble (see Table 1). In the local-scramble condition, adjacent word pair order is preserved, but the order of words within a bigram is swapped. In the global-scramble, words within a sentence are scrambled at random, destroying all structural information. The level of degradedness is reflected in the perplexity on attested English data: 617 for local-scramble, 3161 for global-scramble, but only 104 for control.

The LMs are trained on 13M tokens for 1.2k training steps. We estimate word surprisal from all three LMs and fit linear mixed-effects regression models on three corpora of eye-tracking-while-reading (Dundee [6], Provo [7], GECO [1]) and a corpus of self-paced reading times (Natural Stories [2]). Regression models include word length, log word frequency, and word position, with surprisal added incrementally. We compare model fit with a likelihood ratio test (change in log-likelihood, ΔLL), testing whether surprisal improves the fit.

As shown in Figure 1, our results on first-pass reading time (FPRT), total reading time (TRT) and SPR time reveal that surprisal from the control and global-scramble models improves the fit on all four corpora, with the control one improving it the most. On Dundee, GECO (FPRT) and Natural Stories, the control LM is followed by local-scramble. However, on GECO (TRT) and Provo (both measures), the local-scramble LM does not significantly contribute to RT prediction beyond baseline predictors. The results on Provo diverge from other corpora, possibly due to shorter sentence length relative to others.

Both scrambled LMs lack hierarchical information, but in spite of this they can estimate processing difficulty in some cases. Perhaps the most unexpected finding is that the global-scramble model, which is essentially a linearized bag of words, contains information relevant to the reading behavior on all corpora.

Our findings suggest that structural information together with co-occurrence statistics leads to the most predictive surprisal. Crucially, surprisal estimates remain partially informative even when hierarchical structure is absent from training. Our results motivate future work examining what these scrambled models actually learn, and testing them on constructions where local statistics and hierarchical structure make divergent predictions.

Condition	Example
Control	Fine , take it out of the 25 grand that I gave you .
Local	, Fine it take of out 25 the that grand gave I . you
Global	grand that 25 out , the of Fine you gave I it . take

Table 1: Examples of an original training instance (Control) and the two scrambled versions (Local and Global).

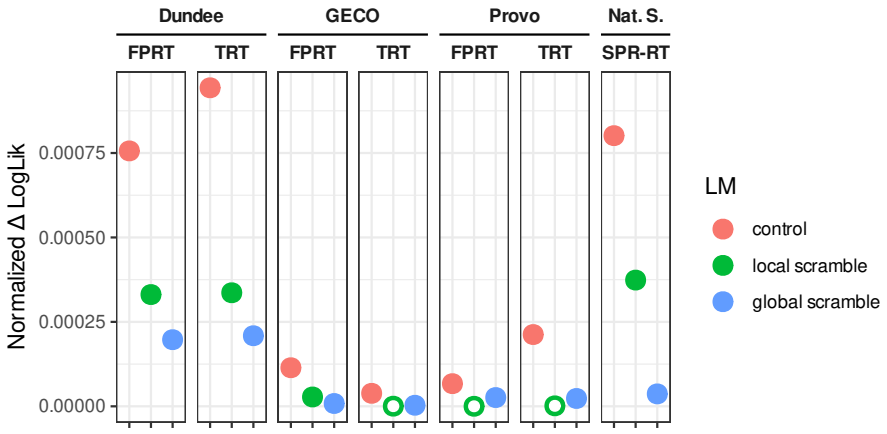

Figure 1: Improvement in fit (measured as ΔLL, normalized given dataset size) with surprisal from three LM conditions. Hollow dots signify a non-significant change in fit ($p > .05$).

References

[1] U. Cop, N. Dirix, D. Drieghe, and W. Duyck. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 2017.

[2] R. Futrell, E. Gibson, H. J. Tily, I. Blank, A. Vishnevetsky, S. T. Piantadosi, and E. Fedorenko. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 2021.

[3] J. Hale. A probabilistic Earley parser as a psycholinguistic model. In *NAACL*. ACL, 2001.

[4] K.-J. Huang, S. Arehalli, M. Kugemoto, C. Muxica, G. Prasad, B. Dillon, and T. Linzen. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 2024.

[5] J. Kallini, I. Papadimitriou, R. Futrell, K. Mahowald, and C. Potts. Mission: Impossible language models. In *Proceedings of ACL*, August 2024.

[6] A. Kennedy, J. Pynte, and R. Hill. The Dundee corpus. In *Proceedings of ECEM*, 2003.

[7] S. G. Luke and K. Christianson. The Provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 2018.

[8] C. Shain, C. Meister, T. Pimentel, R. Cotterell, and R. Levy. Large-scale evidence for logarithmic effects of word predictability on reading time. *PNAS*, 2024.

[9] X. Yang, T. Aoyama, Y. Yao, and E. Wilcox. Anything goes? A crosslinguistic study of (im)possible language learning in LMs. In *Proceedings of ACL*, July 2025.