Large-scale eye-tracking while reading benchmark shows surprisal captures early fixations, but not regressions

William Timkey[1], Suhas Arehalli[2], Kuan-Jung Huang[3], Grusha Prasad[4], Tal Linzen[1], Brian Dillon[3]
[1]New York University, [2]Macalester College, [3]U. of Massachusetts Amherst, [4]Colgate University

Surprisal-based accounts of syntactic disambiguation difficulty hypothesize that incremental processing difficulty in garden path (GP) sentences can be explained by word-level predictability [1]. Contrary to this hypothesis, recent large-scale self-paced reading (SPR) studies found that surprisal from neural language models drastically underpredicts the magnitude of reading time (RT) slowdowns in GP constructions [2,3]. One reason for this failure could be that reading measures in SPR conflate distinct sources of difficulty such as word recognition and syntactic integration into a single RT measure per word. In eye-tracking, on the other hand, different sources of difficulty tend to be associated with distinct measures [4]; for example, word recognition is associated with *first pass* (a.k.a. *gaze duration*; the total fixation time on a word before exiting to the left or right) whereas syntactic integration is associated with *first pass regression out probability* (*RO*; whether the reader regresses to a previous word after the first pass). We conduct a large-scale (N=368) eye-tracking experiment, using the same stimuli and design as [2], to investigate the extent to which surprisal can explain gaze duration and RO in GP constructions.

Methods: Each participant saw 4 sentences of 13 experimental constructions intermixed with 40 filler sentences. Here we focus on only the GP subset (3 GPs x 2 ambiguity, see (1)). Participants answered a comprehension question following each sentence. Participants with an accuracy below 80% on filler questions or with more than 25% track loss/blinking during first-pass of the target word were excluded (13%). Trials with gaze duration on the target word longer than 2 seconds were removed (< 0.1%).

We estimate empirical GPEs by fitting Bayesian mixed-effects regression models to the two reading measures (see model formula in (3)). To generate predicted GPEs from surprisal estimates, we follow the method of [2, 3]: First, we obtained surprisal estimates from two neural language models: GPT-2 and an LSTM [4]. Second, we estimated coefficients predicting gaze duration and RO from surprisal using the filler sentences (while controlling for word length, position, frequency and spillover). Then, we use the conversion factors to generate predicted gaze durations and RO probabilities on the target GP sentences. Finally, we fit Bayesian models (same structure as (3)) to the predicted data to estimate predicted GPEs. We look at two positions: the disambiguating word and the spillover word.

Gaze Duration: The GPE in gaze duration was localized to the disambiguating verb. As in the SPR study, the no-surprisal baseline models did not capture the magnitude of any GPEs. Unlike SPR, surprisal from GPT-2 captures GPE magnitude in two out of three constructions. Surprisal from the LSTM captures one of three. Even though surprisal under-predicted GPEs in the MV/RR construction by a factor of ~4.5, this under-prediction is far less severe than previously observed in SPR (28x). Our results are consistent with the claim that surprisal largely captures the magnitude of GPEs for gaze duration.

Regressions Out: We find GPEs at both the disambiguating verb and spillover word in RO, as well as significant differences in regression probabilities by construction in the spillover region. Surprisal from GPT-2 and the LSTM does not account for the magnitude of the GPE in either region.

Summary: We find evidence that surprisal can better explain GPEs in gaze duration than in RO. This discrepancy suggests distinct processes involved in syntactic disambiguation [4], with surprisal affecting only some of them. We argue that high regression rates in GPs most likely reflects structural reanalysis [4, 6]. This hypothesis can be further evaluated by comparing predictions from models that implement reanalysis against both ROs and regression-path durations across the three GP constructions [e.g. 7, 8].

References

[1] Hale, J. (2001). *NAACL* [2] Huang K.J. et al. (2024) *Journal of Memory and Language*. [3] van Schijndel, M. & Linzen, T. (2021). *Cognitive Science*. [4] Reichle, E. D. et al. (2009). *Psychonomic Bulletin & Review.* [5] Gulordava, K. et al. (2018). *NAACL-HLT* [6] Frazier, L., & Rayner, K. (1982). *Cognitive Psychology.* [7] Ferreira F. & Henderson J.M., (1998). *Reanalysis in Sentence Processing.* [8] Sturt, P. (1997) *Thesis*.

1a. The little girl (who was) fed the lamb **remained relatively** calm despite having asked for beef. (MV/RR)
1b. The little girl found (that) the lamb **remained relatively** calm despite the absence of its mother. (NP/S)
1c. When the little girl attacked(,) the lamb **remained relatively** calm despite the sudden assault. (NP/Z)

Tab 1: An example of a GP triplet. (1a) has a locally ambiguous verb phrase that can be either a main verb (MV) or a reduced relative clause (RR). (1b) has a locally ambiguous noun phrase that can be either the direct object of the verb or the subject of a sentential complement (S). (1c) has a locally ambiguous noun phrase that can be either the direct object or the subject of an upcoming independent clause. Critical position and the spillover positions in bold. Parentheses denote the unambiguous forms. Example stimuli and their descriptions are adapted from [3]. The present experiment included all constructions and stimuli from [2], but we only investigate GP constructions in the present work.

(2a) Surprisal filler models:

$$Reading\_measure \sim Surp(w_i) + Surp(w_{i+1}) + Pos(w_i) + Freq(w_i)*Len(w_i) + Freq(w_{i+1})*Len(w_{i+1}) + (1 + Surp(w_i) + Surp(w_{i+1})|| subj) + (1 | item)$$

(2b) Baseline filler models:

$$Reading\_measure \sim Pos(w_i) + Freq(w_i)*Len(w_i) + Freq(w_{i+1})*Len(w_{i+1}) + (1 + Freq(w_i) + Freq(w_{i+1})|| subj) + (1 | item)$$

(3) GP models (Bayesian):

$$reading\ measure \sim Ambiguity*(NP/Svs.MV/RR + NP/Vvs.MV/RR) + (1 + Ambiguity * (NP/Svs.MV/RR + NP/Vvs.MV/RR)|subj) + (1 + Ambiguity * (NP/Svs.MV/RR + NP/Vvs.MV/RR)|item)$$

Tab 2: Details about statistical model. Surp(w) = surprisal of word w, Pos(w) = position, Len(w) = length, Freq(w) = log unigram frequency. For the filler models, we use logistic regression to predict RO, a binary variable, and linear regression to predict gaze duration.

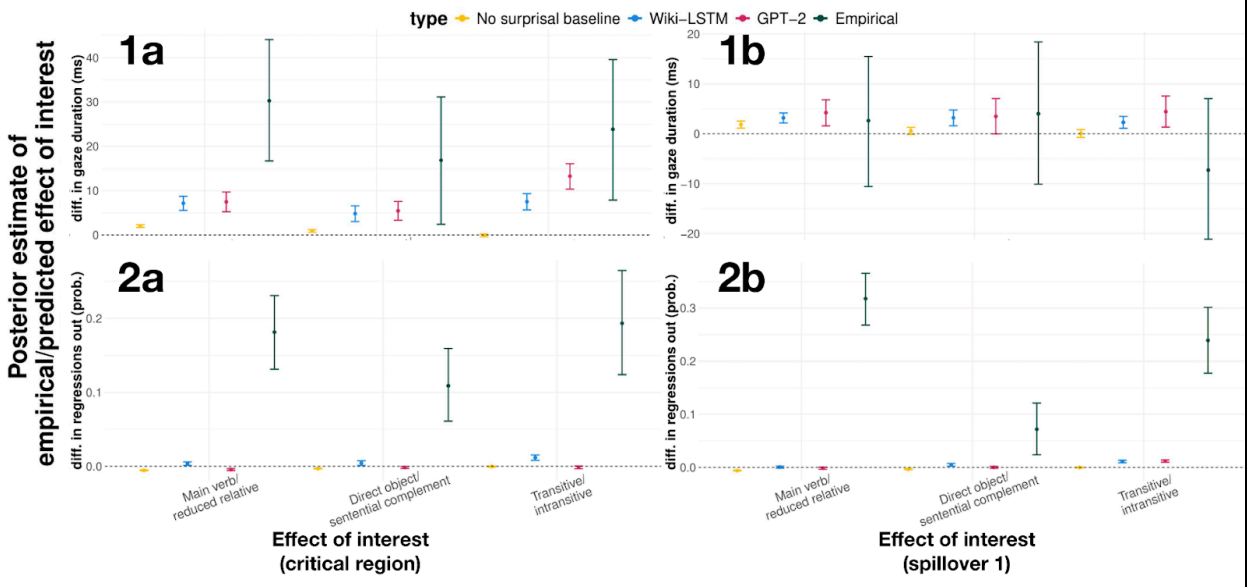

Figure 1: Empirical and predicted GPEs in each construction. Gaze duration at the disambiguating verb is given in (1a) and the spillover region in (1b). RO proportions are given at the disambiguating verb at (1c) and the spillover region in (1d). Error bars represent posterior 95% quantile ranges.