SPR mega-benchmark shows surprisal tracks construction- but not item-level difficulty

Kuan-Jung Huang[1], Suhas Arehalli[2], Mari Kugemoto[1], Christian Muxica[3], Grusha Prasad[2], Brian Dillon[1], Tal Linzen[4]
[1]University of Massachusetts Amherst, [2]Johns Hopkins University, [3]University of California Los Angeles, [4]New York University

The surprisal account of syntactic disambiguation difficulty holds that word-level unpredictability is the sole determinant of processing difficulty in garden path (GP) sentences [1]. This entails that surprisal should predict processing difficulty both across different constructions and across individual items. We test this by looking at construction- and item-level GP effects (GPEs) in a large-scale self-paced-reading (SPR) benchmark we introduce: the Syntactic Ambiguity Processing (SAP) Benchmark, which has orders of magnitude more data than a standard reading experiment. We focus on a subset of the constructions in this dataset: MV/RR (1a), NP/S (1b), and NP/Z (1c)).

The SAP benchmark uses a standard within-item factorial design to estimate GPEs. We created twenty-four sentence triplets as in (1). Each participant saw 4 ambiguous and 4 unambiguous instances of each construction. These sentences were intermixed with other sentence constructions from the benchmark and 30 diverse filler sentences. Participants answered a comprehension question following each sentence; only those whose accuracy on fillers was 80% or higher were analyzed (N=2000; recruited on Prolific). There were 220–440 datapoints per item, yielding more precise item-level estimates of GPEs compared to prior work.

First, to empirically probe the relative difficulty among the GP constructions, we ran three Bayesian maximally-structured mixed-effect models, each for the disambiguating verb, the first spillover word and the second spillover word. At the disambiguating verb, NP/Z had the largest GPE, followed by MV/RR and then NP/S (Fig. 1a). For all constructions, GPEs peaked at the first spillover position, but the peak was much higher in the MV/RR construction compared to the others. Second, to estimate item-level GPEs, we used another Bayesian LMM, following [2].

Next, we tested whether surprisal accounts for the observed GPEs, by deriving surprisal estimates from two language models (LMs, models that output next-word probability): a Transformer LM (GPT-2) and an LSTM trained on an 80M word subset of English Wikipedia (Wiki-LSTM). For each item, we subtracted the surprisal estimates of the disambiguating verb in the ambiguous and unambiguous conditions. At the construction level, both LMs correctly predicted the relative difficulty at the disambiguating verb (Fig. 1b). We then correlated within each construction, item-wise, the surprisal differences and the empirical GPEs. Wiki-LSTM failed to explain variance within NP/Z, and GPT-2 did not explain any variance within NP/S or within MV/RR (Fig. 2). This suggests that when we use next-word probabilities from language models to operationalize surprisal theory, surprisal alone cannot account for item-level processing difficulty. This discrepancy cannot be attributed to a failure of these neural models to capture information beyond corpus statistics: prior work has shown that these models capture semantic and thematic relations between words [4,5] and are sensitive to syntactic structure [3].

To summarize, first, we found an alignment between GPEs and surprisal differences at the construction level. This contrasts with [3], likely reflecting our much larger sample size and additional controls. Meanwhile, substantial item-wise variation existed within constructions, and it was not well captured by the LMs. Finally, the drastic increase in the GPE at the first spillover position for MV/RR suggests either a delay in noticing disambiguation or additional downstream (re)processing for this construction [6], though further formal analyses controlling for spillover effects in SPR are needed [3]. This observation, along with the inability of the LMs to capture item-level processing difficulty, suggest that GPEs may not be reducible to surprisal alone.

References

[1] Hale, J. (2001). *NAACL* [2] Rouder, J. & Haaf, J. (2019). *Psychonomic Bulletin & Review* [3] van Schijndel, M. & Linzen, T. (2021). *Cognitive Science* [4] Frank, S. & Hoeks, J. (2019). *Proceedings of the Cognitive Science Society* [5] Michaelov, J. A., Bardolpha, M. D., Coulson, S. & Bergen, B. K. (2021). *Proceedings of the Cognitive Science Society* [6] Fodor, J. & Ferreira, F. (1998). *Springer Science & Business Media*

(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)
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 two spillover positions in bold. Parentheses denote the unambiguous forms.

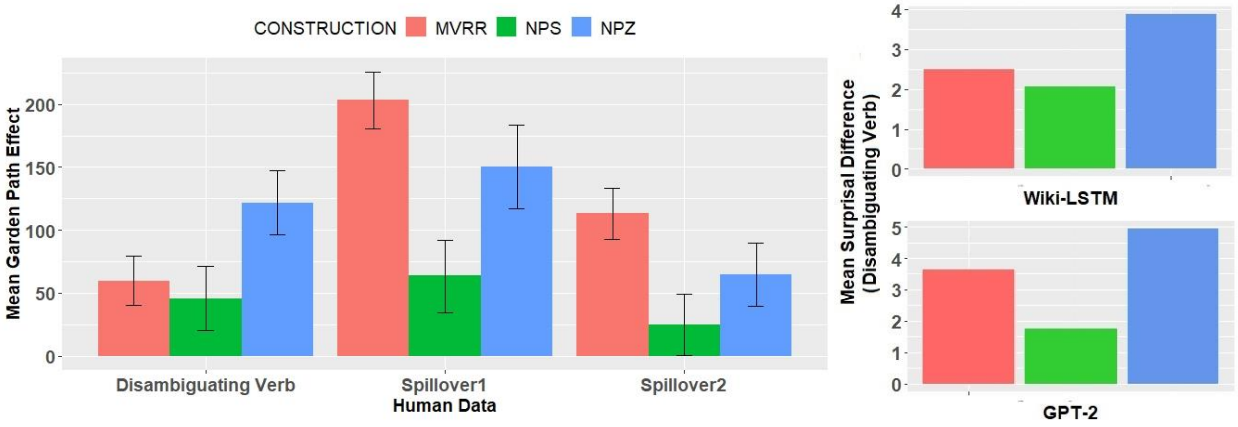

Figure 1. (1a) Empirical mean GPEs for each construction at the three critical words. Error bars reflect 95% quantile ranges of the posteriors. Models: RT~ambiguity*construction+(1+ambiguity*construction||item)+(1+ambiguity*construction||subj). Uninformative priors used. Warmup = 500; Iter = 4000; Chain = 4. All Rhats <= 1.01.
(1b) Mean surprisal difference estimated from the two LMs. Note that, with our current model specification, surprisal is not expected to predict GPEs at spillover regions.

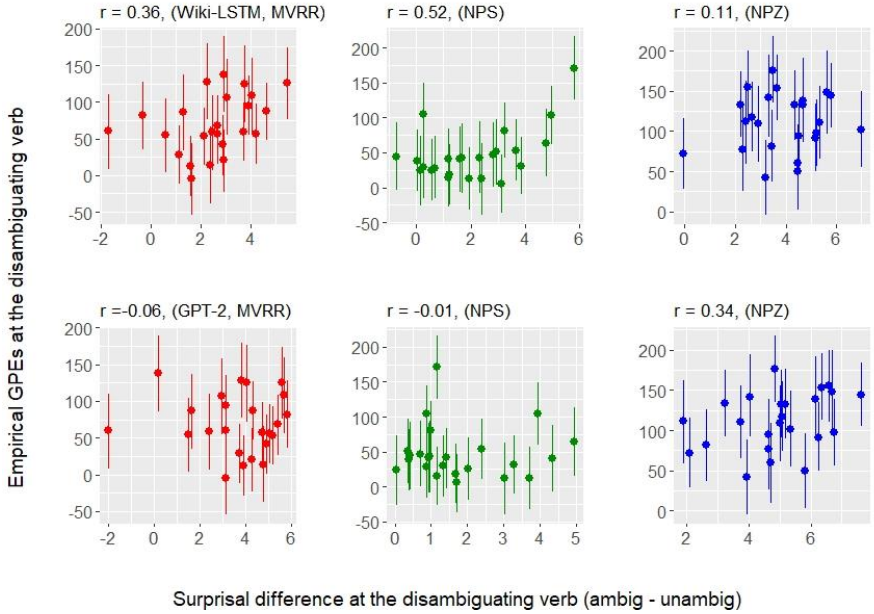

Figure 2. Scatterplots of item-wise surprisal differences and GPEs at the verb. Error bars reflect 95% quantile ranges of the posteriors. Note the ranges of the y-axes revealed that even with only 24 items per construction, there was substantial variation in the magnitude of GPEs.