LM surprisal underpredicts garden path effects even with limited syntactic parallelism

William Timkey[1], Tal Linzen[1] [1]New York University

Surprisal-based accounts of syntactic processing [1,2] posit that the magnitude of *garden path effects* (GPEs) can be explained by the surprisal (negative log probability) of the disambiguating word in context. Recent tests of this hypothesis have found that surprisal estimates from neural language models (LMs) drastically underestimate the magnitude of GPEs in human eyetracking [3] and self-paced reading experiments [4,5]. One possible explanation for this underestimation is that LMs trained without an explicit syntactic parsing objective can implicitly maintain beliefs over an unbounded number of syntactic parses in parallel [6], while humans might only be able to consider as few as one interpretation of a sentence at a time. If a sentence is disambiguated in favor of an interpretation that wasn't under consideration in the first place, then the disambiguation point should be highly surprising, leading to a large GPE. If LMs implicitly assign too much probability to the globally correct parse, then the disambiguating word should be relatively unsurprising, leading to small GPEs. In this work, we explore the hypothesis that LMs constrained to consider fewer parses at a time are better at capturing human GPE magnitudes.

Models: We use a class of LMs whose syntactic parallelism can be explicitly manipulated (Recurrent Neural Network Grammars (RNNGs) [7] with Word-Synchronous Beam Search (WSBS) [8,9]). RNNGs are trained to jointly predict upcoming words and parsing actions. We train 5 random seeds of an RNNG on a machine-parsed version of the 60m token BLLIP corpus, using a left-corner parsing strategy. Trained RNNGs can be combined with the WSBS inference algorithm to incrementally search for candidate parses given a sequence of words, and generate word-level surprisal estimates. The *word beam width* (k_w) parameter of WSBS determines how many distinct syntactic parses the model will consider at each word. WSBS calculates a word-level surprisal by marginalizing over all k_w parses. We obtain surprisal estimates for experimental items using $k_w \in \{1,2,3,4,5,10,25,50,100,250,500,1000\}$.

Methods: We use materials and self-paced reading data from [4]. In Experiment 1, we compute GPEs, measured in bits of surprisal, for each syntactic construction and beam size. In Experiment 2, we predict GPE magnitudes in reading times by first fitting linear mixed-effects regression (LMER) models to predict reading times from surprisal on experimental filler sentences (while controlling for word length, position and frequency), then predicting reading times from surprisal on GP sentences using the fitted models, using an identical LMER model structure to [5]. We also analyse the predictive power of each filler model on the filler sentences to see whether there is a mismatch between k_w that best predicts GPEs and reading times in filler sentences. In Experiment 3, we compute a loose upper-bound on GPE size by forcing models to only consider the initially preferred, but ultimately incorrect parse of the sentence. The disambiguating word is an ungrammatical continuation of the initially preferred interpretation, and should therefore have high surprisal. We manually remove all instances of the globally correct parse from the beam in the ambiguous condition, leave the unambiguous condition unchanged, and calculate the resulting surprisals and GPEs.

Results: In Experiment 1, we find that GPEs are generally larger for models with more restricted parallelism when measured in surprisal (lower k_w; Figure 1, top). While Exp. 1 generally supports our hypothesis, Experiment 2 shows that GPE magnitudes are still drastically underpredicted across all k_w, when measured in predicted reading times. (Figure 1, bottom). On filler sentences, the predictive power of the RNNGs peaks around $k_w = 50$, which is generally higher than the k_w which produced the largest GPEs ($1 < k_w < 5$) (Figure 2). Experiment 3 shows that RNNGs underpredict GPEs even when models are forced to only consider the ultimately incorrect interpretation of the sentence (Figure 1, red bars), meaning the LMs failures to capture GPE magnitudes are not strictly driven by excessive syntactic parallelism, but rather by a failure in RNNGs to assign sufficiently low probabilities to ungrammatical continuations. Our results suggest that GPEs might be better captured by LMs whose predictions are better calibrated to grammatical constraints, or models with limited parallelism and explicit reanalysis mechanisms.

References: [1] Hale, J. (2001). *NAACL* [2] Levy, R. (2008) *Cognition* [3] Timkey et al. 2024 *HSP* [4] Huang K.J. et al. (2024) *JML*. [5] van Schijndel, M. & Linzen, T. (2021). *Cognitive Science*. [6] Jurafsky (1996) *Cognition* [7] Dyer et al. 2016 *NAACL* [8] Stern et al. (2017) *EMNLP* [9] Hale et al. (2018) *ACL* [10] Hu et al. (2020) *ACL*

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)**

GPE: RT("remained" | "The little girl fed the lamb") - RT("remained" | "The little girl who was fed the lamb")

Table 1: An example of a GP triplet from the SAP benchmark dataset [4]. Colors denote the critical word, spillover 1, and spillover 2. Parentheses denote material only present in the unambiguous conditions. (1a) has a locally ambiguous verb phrase that can be either a main verb (MV) or a reduced relative clause modifying the subject (RR). (1b) has a locally ambiguous noun phrase that can be either the direct object complement of the verb or the subject of a sentential complement of the verb (S). (1c) has a locally ambiguous noun phrase that can be either the direct object complement of the verb, or the subject of an upcoming independent clause. GPEs are calculated as the difference in reading times at the regions of interest across the ambiguous and unambiguous conditions.

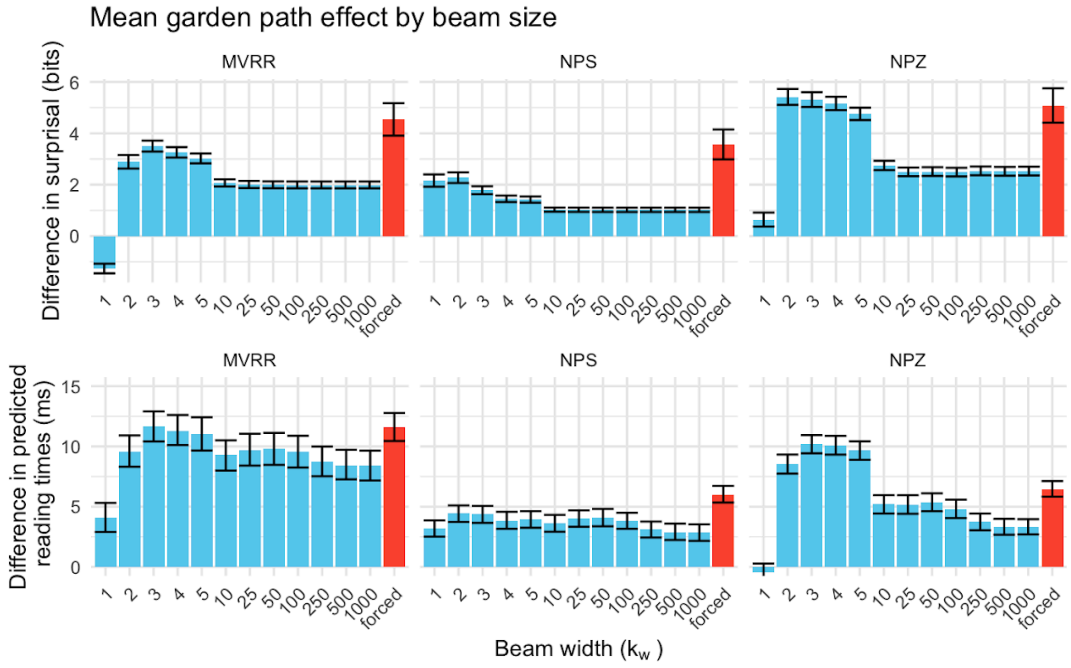

Figure 1: Mean GPEs by WBBS word beam size (k_w), averaged across items, the three regions of interest, and the 5 RNNG model random seeds, measured in surprisal (top) and predicted reading times (bottom). Red bars labeled "forced" summarize the results of Exp. 3, which measures the predicted GPEs at k_w = 1000 when all parses consistent with the globally correct interpretation of the sentence are manually pruned from the beam in the ambiguous condition. Error bars represent bootstrapped 95% confidence intervals. (Empirical GPE magnitudes in humans reported in [4] are 125ms for MVRR, 45ms for NPS, and 112ms for NPZ)

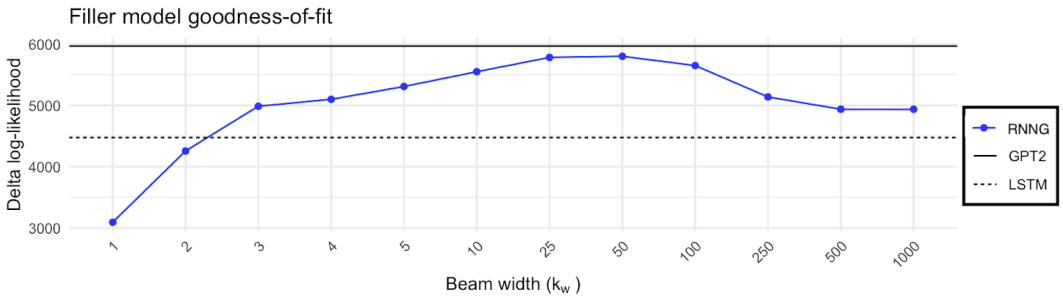

Figure 2: Goodness-of-fit of filler models at various beam sizes, measures as the delta log likelihood of LMER models with and without surprisal-based predictors. LSTM and GPT-2 results from [4] are included for reference.