Prediction and structural processing dissociate: A large-scale eyetracking study

William Timkey[1], Kuan-Jung Huang[2], Byung-Doh[3] Oh, Grusha Prasad[4], Suhas Arehalli[5],Tal Linzen[1], Brian Dillon[6]

[1]NYU, [2]MIT/UMD, [3]NTU Singapore,[4]Colgate U. [5]Macalester College, [6]UMass Amherst

What are the linguistic computations that drive eye movements in reading? Under one view (*surprisal theory* [1,2]) the fundamental work of comprehension is prediction: the total cost of processing each word in a sentence is fully reducible to its surprisal. An alternative view [e.g. 3,4] holds that comprehension reflects the cost of structure-building operations that cannot be reduced to predictability alone. Under this view, both prediction and structural processing could affect eye movements in distinct, dissociable ways (e.g.. at different stages of processing). Earlier attempts to address this debate have been inconclusive because of small numbers of participants, paradigms that produce only a single word level reading measure (e.g. self-paced reading [7,8,10] and maze [9]), and a limited set of predictability estimators [7-10]. We conduct a large-scale study (*n*=368) examining eye movements in garden path constructions (GP), and object vs. subject relative clause processing asymmetries (RC). We compare human reading behavior to surprisal estimates from 407 neural language models (LMs) across 6 architectures and various training settings.

Methods: Materials were identical to [7]. Each participant saw 4 sentences of 6 experimental constructions, each with two conditions, intermixed with 40 filler sentences. Here we focus on 4 of these: 3 GP constructions and the RC construction. We analyze 4 reading measures in total: (1) *Forward reading time*: The total duration of all first-pass fixations on a word in trials where the word is exited to the right. (2) *Regressive gaze duration*: The total duration of all first-pass fixations on a word in trials where the word is exited to the left. (3) *Regressive go past*: the sum of the duration of all first-pass fixations on a word in trials where the word is exited to the left, plus the duration of all subsequent fixations on any preceding words before it is exited to the right. These 3 measures are *regression-contingent* and were chosen to tease apart the known influence of regressions on the duration of first-pass reading [11]. (4) *First pass regressions out*: The proportion of trials a word is exited to the left during the first pass. We estimate processing effects by fitting Bayesian mixed-effects regression models to the 4 reading measures with identical formulae to [7]. We follow the established methodology of [8] to generate surprisal theory-predicted reading measurements for each of the 4 measures from each of the LMs. This method predicts reading times in GP/RC sentences from their surprisals based on the RT~Surprisal relationship observed in *filler* sentences. Under surprisal theory, the RT~Surprisal relationship in filler sentences should predict the magnitude of garden path and relative clause processing effects. We estimate effects from the surprisal-predicted RTs using regression models with identical specifications to the empirical ones, and compare the magnitude and direction of the surprisal-predicted effects to the empirical effects.

Results: In contrast to prior SPR results, where surprisal uniformly fails to explain garden path effects, we find that surprisal *can* explain the direction and magnitude of most GP effects, but only in forward reading time. Effects in measures that index regressive reading (regressive gaze, regressive go-past, regressions out) are drastically underpredicted by surprisal. The dissociation between forward and regressive reading was stable across LM families; LMs with explicit syntactic representations (RNNGs) did not outperform standard architectures, meaning the failure of LM-based surprisal estimators to explain rereading behavior is likely not due to low-quality syntactic representations. This distinction aligns with the predictions of theories of oculomotor control in reading [e.g. 5]: FRT is argued to index word recognition [5], while regressions are associated with failures in integrating a recognized word with its context [6]. Our results suggest that LM surprisal is sufficient to explain probabilistic influences on word recognition (a process active in simple and complex sentences alike) but cannot explain the frequency nor time cost of integration failures in syntactically challenging sentences. In order to explain regressive reading behavior, future cognitive models should pair predictive processes with equally precise accounts of how readers construct and revise syntactic representations during real-time processing.

References [1] Hale, J. (2001). *NAACL* [2] Levy, R. (2008). *Cognition* [3] Frazier, L. & Fodor J. (1978) *Cognition*. [4] Lewis, R. & Vasishth, S. (2005) *Cognitive Science* [5] Reichle, E. D. et al. (2009). *Psychonomic Bulletin & Review.* [6] Staub, A. (2011) *Journal of Experimental Psychology* [7] Huang K.J. et al. (2024) *Journal of Memory and Language*. [8] van Schijndel, M. & Linzen, T. (2021). *Cognitive Science*. [9] Wilcox, E. et al. (2021) *ACL-IJCNLP 2021* [10] Kobzeva, A. & Kush D. (2024). *Cognitive Science* [11] Altmann, G. et al. (1992) *JML*.

(1a) Surprisal filler models:
*Reading_measure ~ 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)*

(1b) Baseline filler models:
*Reading_measure ~ 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)*

(2) Garden Path models (Bayesian):
reading measure ~ 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)*

(3) Word position residualization model for relative clause subset (fit to filler sentences):
reading_measure ~ Position(wi) + (1 + Position(wi) || subject) + (1 | item)
(measure_corrected = actual_reading_measure - predicted_reading_measure)

(4) Garden Path models (Bayesian):
measure_corrected ~ SRC_vs_ORC + (0 + SRC_vs_ORC || subject) + (1 + SRC_vs_ORC || item)

Table 1: 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 the other measures.

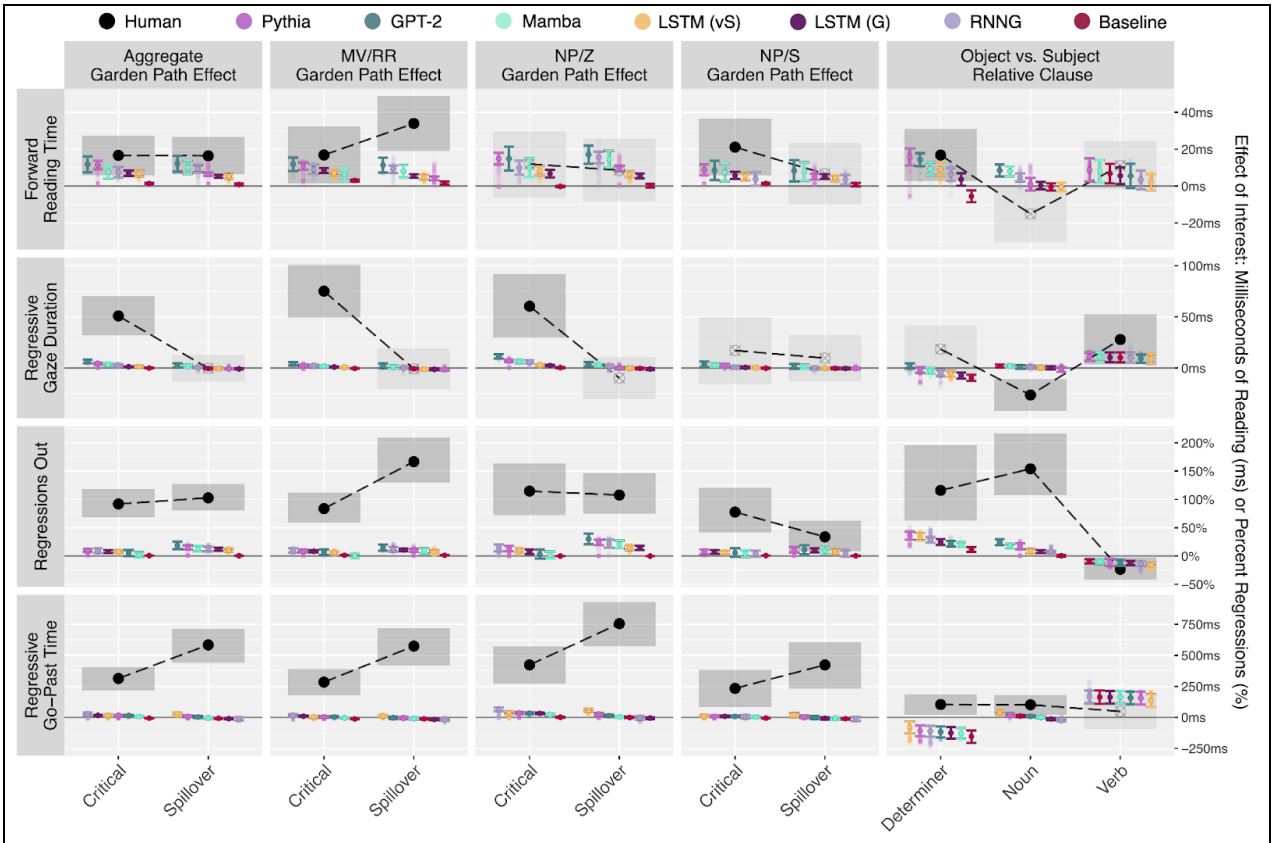

Figure 1: Empirical (Human) and predicted (LM) effects of interest at the regions of interest in four selected constructions (columns) and four measures (rows). Human effect estimates are represented with black points for the mean, with shaded regions representing 95% credible intervals in the vertical dimension. Empirical null effects are denoted with lighter shading and an X on the mean estimate point. Colored points are predicted effect sizes from the LM within each architecture whose surprisal values provided the best fit to reading data in the filler sentences, measured by filler model ΔLL. Within each plot, predicted LM effects are sorted from largest to smallest, with error bars denoting 95% credible intervals, incorporating participant-level and item-level uncertainty. Mean effect estimates from all 407 LLM-derived surprisal estimates are displayed as square semi-transparent colored points without error bars.