Surprise! LM-generated Surprisal Values Cannot Fully Account for Human Parsers' Preferences in Relative Clause Comprehension

Liliana Nentcheva, Sebastian Schuster & Andrea Santi, University College London

It remains an open question the extent to which next-word prediction can account for human sentence processing behaviour. One extreme view is that processing difficulty can be entirely reduced to the probabilities which a parser assigns to each word in the context of the preceding string and whether these align with material in the input [1,2]. This position advocates the success of computerised language models (LMs) in making highly accurate next-word predictions and the strong correlation this has with an LM's ability to predict human neural and behavioural responses during the processing of (simple) sentences [2]. However, recent data demonstrates that LM-generated next-word probabilities severely underestimate human processing difficulty with complicated garden-path structures [3]. Purely expectation-based theories may therefore have limited ability to explain processing difficulty with complex sentences, particularly those that incur high memory costs [3] or benefit from hierarchical structure for their interpretation [4].

In extending this work beyond processing difficulty, we ask if LM-generated next-word probabilities directly align with parsing preferences in local and non-local dependencies, in English. We use data from a modified version of the Maze Task [5], in which participants were required to choose between two possible continuations of a relative clause (RC) following a choice point: one continuation was compatible with a subject gap (e.g. *should*) and the other with an object gap (*the*) (**Table 1**). The choice point was either non-local (Experiment 1, n=76) following an RC verb that selected for a complement clause, or local (Exp. 2, n=62) appearing immediately after the relativiser. Human data showed a preference for subject-gap continuations across both experiments, although their probability was increased when the decision point was non-local [5].

We subsequently generated estimates of a word's negative log-probability in context, i.e. Surprisal [1], for the subject and object gap continuations at the decision point (*should* vs. *the*) using GPT-2 (small): the LM that best predicts human processing difficulty [6]. The Surprisal values were converted into the probability of each parse (subject vs. object gap) and used to calculate the relative probability of a subject gap ($\mathrm{Sgap}_{\mathrm{surprisal}}$) (**Equation 1**). Using human and machine data on the same scale avoids the assumptions of a linking function, as has previously been used [2,3]. A logistic regression showed an interaction between $\mathrm{Sgap}_{\mathrm{surprisal}}$ and locality ($p < .001$): $\mathrm{Sgap}_{\mathrm{surprisal}}$ predicted the subject-gap preference locally, but failed to non-locally (**Tables 2** and **3**). In the non-local case, human preference data showed a strong subject-gap preference (0.91) significantly exceeding that of LMs (0.65) (**Figure 1**).

The results confirm that LM-generated Surprisal has limited explanatory power in the context of complex sentences even when reanalysis is not required, as with the non-local parsing preferences in [5]. Potential explanations include constraints on humans' memory resources, which are not accurately represented in the LM, and may contribute to a differential weighting of lexical vs. syntactic factors in next-word prediction [3]. Specifically, increased memory demands in the non-local dependency could engage mechanisms such as the Active Filler Strategy [7] which encourage the parser to offload the filler early and could take priority over frequency considerations. Future work should explore what adjustments can be made to the LMs so that they better approximate human parsing preferences, which will enable us to better understand the predictive mechanisms engaged in human sentence processing.

Table 1. *Example items;* ***bolded*** *word indicates where Surprisal and parser preferences were measured in RCs with (Exp. 1) and without (Exp. 2) an intervening clause*

Experiment	*Pre-Decision Point*	*Subject/Object Gap Continuation*	*Matrix Clause*
1. Non-local	*The girl* who the teacher remarked*	***should** be rewarded by the department*	*had surpassed all expectations*
		***the** department should reward*	
2. Local	*The teacher remarked that the girl who*	***should** be rewarded by the department*	
		***the** department should reward*	

*Filler animacy was also manipulated [5], which does not bare on the current hypotheses or interpretations, but it is included as an explanatory factor in the statistical models.

Equation 1. *Probability of a subject gap parse (pSgap) calculated from Surprisal values*

(i) $\text{Surprisal}_{subj} = -\log P(w_{det}|1.....w_{det-1})$; $\text{Surprisal}_{obj} = -\log P(w_{aux}|1....w_{aux-1})$

(ii) $P(w_{det}|w1...w_{det-1}) = 2^{-\text{Surprisalsubj}}$; $P(w_{aux}|w1...w_{aux-1}) = 2^{-\text{Surprisalobj}}$

(iii) pSgap: $\text{pSgap} = 2^{-\text{Surprisalsubj}}/(2^{-\text{Surprisalsubj}}+2^{-\text{Surprisalobj}})$

Table 2. Results *based on the full model* *(R^2_m=0.23; R^2_c=0.44): glmer(Continuation ~ Animacy * Experiment * pSgap + (1+Animacy|Participant)+(1+Animacy|Item)*

Fixed Effects	β	*Std. Error*	*z-value*	*p-value*
Animacy	0.021	0.208	0.100	.920
Experiment	1.916	0.239	8.011	< .001
pSgap	0.901	0.305	2.954	.003
Ani*Exp	0.167	0.207	0.808	.419
Ani*pSgap	0.403	0.287	1.407	.159
Exp*pSgap	-1.281	0.311	-4.116	< .001
Ani*Exp*pSgap	-0.033	0.281	-0.119	.905

Figure 1. *Human vs Suprisal: Probability of a Subject Gap by Dependency Length*

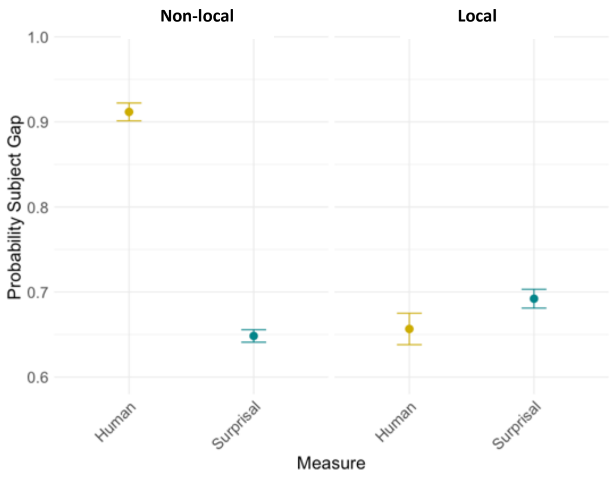

Table 3. *Results based on the simple effects model: glmer(Continuation ~ Animacy + Experiment/pSgap +(1+Animacy|Participant)+(1+Animacy|Item)*

Fixed Effects	β	*Std. Error*	*z-value*	*p-value*
Animacy	0.284	0.077	3.676	< .001
Experiment	1.727	0.225	7.670	< .001
Exp. 1: pSgap	-0.388	0.501	-0.775	0.438
Exp. 2: pSgap	1.815	0.332	5.460	< .001

References

[1] Levy, R. (2008) Expectation-based syntactic comprehension. Cognition, 106(3)

[2] Schrimpf, M. ... (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45)

[3] Huang, K. … (2024) Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty, *J of Mem and Lang.*, 137

[4] Wolfman, M. … (2024). Hierarchical syntactic structure in human-like language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics 2024*

[5] Nentcheva, L. & Santi, A. (2024). *Parsers Predict Subject Gaps Even for Inanimate Fillers* Poster at AMLaP 2024

[6] Oh, B. & Schuler, W. (2023). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*; 11

[7] Frazier, L., & Clifton, C. (1989). Successive cyclicity in the grammar and the parser. *Language and Cognitive Processes*, 4(2)