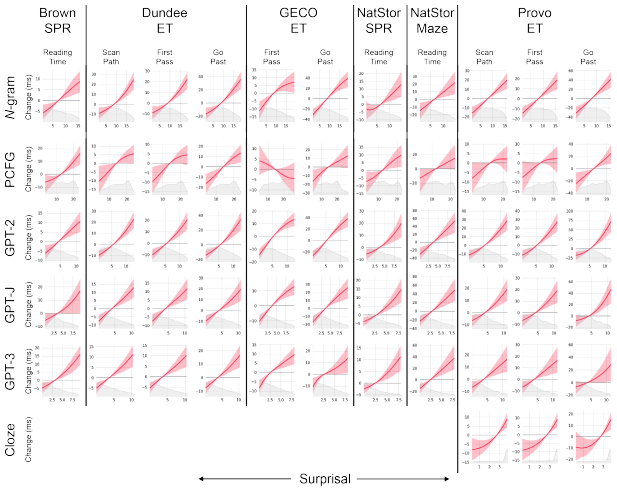
Large-Scale Evidence for Logarithmic Effects of Word Predictability on Reading Time

C. Shain[1], C. Meister[2], T. Pimentel[3], R. Cotterell[2], R. Levy[1] [1]MIT, [2]ETH, [3]Cambridge

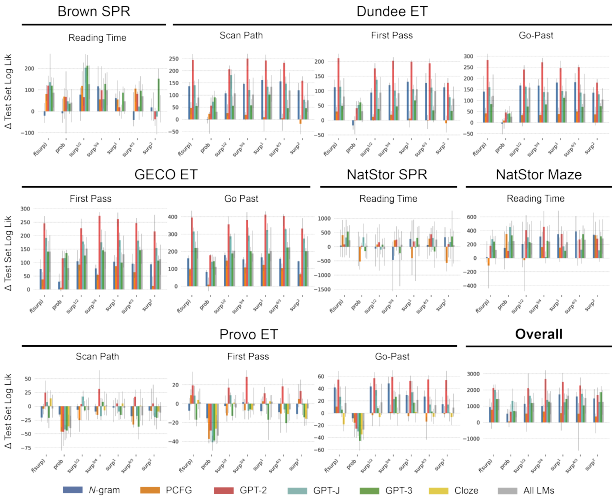
The difficulty of processing a word is related to its predictability in context [19, 5, 17, 2], but why do predictability effects exist? In one view, predictability effects reflect *anticipatory facilitation* of structure-building operations in working memory when words are highly predictable from context [2]. In another view, predictability effects reflect the *cost of probabilistic inference* over a vast space of possible sentence interpretations [7, 11]. If predictability effects reflect facilitation, they should be primarily driven by predictable words, since little advance processing can be done when predictability is low. Indeed, a linear facilitation on next-word predictability falls out from either single-guessing or proportional preactivation prediction strategies [17, 2]. By contrast, if predictability effects reflect the costs of probabilistic inference, they should primarily be driven by unpredictable words, since these words convey more information about how to update the interpretation distribution. Such theories predict either a logarithmic [17] or superlogarithmic [9] (i.e., rapidly increasing) cost as predictability decreases, depending on whether these costs are hypothesized to be linked to the *surprisal* (negative log probability) of a word (logarithmic), or to surprisal plus additional pressures favoring uniform information density (UID, superlogarithmic). Evidence is currently mixed, with studies supporting linear [2], logarithmic [17], and superlogarithmic [9] predictability effects, or some combination of the above [18]. Here we revisit this question at scale by analyzing 6 naturalistic English reading datasets [10, 17, 4, 12, 6, 1], estimating surprisal using diverse language models [8, 20, 15, 21, 3], and applying advanced nonlinear regression techniques [16].

As shown in **Fig 1a**, across datasets and language models, regressions find a largely logarithmic predictability (linear surprisal) effect. To go beyond visualization, we compare the likelihood assigned by each regression model to an unseen test set, relative to a baseline model containing no predictability measure. Results plotted in **Fig 1b** show the performance of models with no constraints on the estimated predictability-cost function ($f(\text{SURP})$) compared to models that are constrained to be linear on probability (PROB) or some fixed exponent of surprisal ($\text{SURP}^{1/2}$, $\text{SURP}^{3/4}$, SURP^{1}, $\text{SURP}^{4/3}$, SURP^{2}). In comparisons across all datasets (bottom right), (i) both the $f(\text{SURP})$ models (which generally find logarithmic effects, **Fig 1a**) and the strictly logarithmic SURP^{1} models substantially and significantly outperform models that are linear on predictability (PROB), and (ii) unconstrained $f(\text{SURP})$ models do not improve on strictly logarithmic SURP^{1} models. We also find that logarithmic (SURP^{1}) or slightly sublogarithmic ($\text{SURP}^{3/4}$) models significantly outperform superlogarithmic models ($\text{SURP}^{4/3}$, SURP^{2}), contrary to the predictions of UID (cf., e.g., [13, 9]; slightly sublogarithmic effects are not predicted by any current theory of language processing). We also highlight two important incidental findings. *First*, GPT-2(-small) substantially and significantly outperforms GPT-3 as a model of human reading, even though GPT-3 has 1000x more parameters, is trained on more data, and has lower perplexity. This suggests that very large transformer language models may be super-human at next-word prediction, which may harm their psychometric performance. *Second*, GPT-2 substantially and significantly outperforms human cloze estimates in the only dataset that provides them (Provo [12]), in line with recent findings [14, 18], suggesting that advanced statistical language models may be preferable to cloze as estimators of human subjective surprisal.

In conclusion, results using large datasets and advanced modeling primarily support a logarithmic effect of word predictability [17], such that small absolute differences in low probability translate to large differences in processing cost, as indexed by reading time. This outcome favors an interpretation of predictability effects as primarily reflecting the costs of probabilistic inference, rather than facilitation at highly predictable words. Results also yield evidence against the superlogarithmic effects predicted by UID, and thus do not support the hypothesis that comprehension processes favor uniform information density.



(a) CDRNN-estimated functional form of surprisal (x-axis) effects on reading times (y-axis) across language model types (n-gram, PCFG, GPT-2, GPT-J, GPT-3, and human cloze) with no delay (i.e. at the surprising word). Kernel density plots show the distribution of surprisal values in the training data over the plotted range.



(b) Change in test set log likelihood as a function of (i) language model and (ii) predictability-cost function over a baseline model containing no predictability measure. Error bars show standard deviation across the ensemble of 10 CDRNN models.

References

[1] Boyce, V. and Levy, R. P. 2022.

[2] Brothers, T. and Kuperberg, G. R. *Journal of Memory and Language*, 2021.

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., and Amodei, D. In *Proceedings of Advances in Neural Information Processing Systems 33*, 2020.

[4] Cop, U., Dirix, N., Drieghe, D., and Duyck, W. *Behavior research methods*, 2017.

[5] Frank, S. L. and Bod, R. *Psychological Science*, 2011.

[6] Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., and Fedorenko, E. *Language Resources and Evaluation*, 2020.

[7] Hale, J. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics*, 2001.

[8] Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 8 2013.

[9] Hoover, J. L., Sonderegger, M., Piantadosi, S. T., and O'Donnell, T. J. *PsyArXiv*, 2022.

[10] Kennedy, A. and Pynte, J. *Vision Research*, 2005.

[11] Levy, R. *Cognition*, 2008.

[12] Luke, S. G. and Christianson, K. *Behavior research methods*, 2018.

[13] Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., and Levy, R. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[14] Oh, B.-D. and Schuler, W. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?, 2022.

[15] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. *OpenAI Blog*, 2019.

[16] Shain, C. and Schuler, W. *arXiv preprint arXiv:2209.12128*, 2022.

[17] Smith, N. J. and Levy, R. *Cognition*, 2013.

[18] Szewczyk, J. M. and Federmeier, K. D. *Journal of Memory and Language*, 2022.

[19] Van Petten, C., Coulson, S., Rubin, S., Plante, E., and Parks, M. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1999.

[20] van Schijndel, M., Exley, A., and Schuler, W. *Topics in Cognitive Science*, 2013.

[21] Wang, B. and Komatsuzaki, A. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, 5 2021.