The Role of Language Model Entropy in Length and Frequency Effects During Reading

Nikki G. Fackler[1] & Peter C. Gordon[1], [1]The University of North Carolina at Chapel Hill

Eye movements during reading provide a window into the online processes behind language comprehension. Two of the most robust predictors of eye movement behavior are word length and word frequency: short words are skipped more frequently and fixated for shorter durations than long words, and the same can be said for more frequent words relative to less frequent words[1]. While context clearly plays a role in word recognition during reading (e.g., predictability effects[2]), word length and frequency are static characteristics of words that have fixed values independent of the word's context. The present study aimed to assess whether the effects of length and frequency on a word are dependent on how constraining or neutral the context is, using information entropy calculated from a large language model (LLM).

Information entropy[3] describes the degree to which an outcome is random; a fair coin flip has high entropy, as the two outcomes (heads or tails) are equally probable. With respect to language processing, entropy describes the degree to which a context is constraining or neutral. Neutral contexts provide little information about what the next word is going to be, whereas constraining contexts bias the likelihood that certain words will occur next. With advancements in artificial intelligence, recent work has utilized next-word probabilities from LLMs to compute candidate measures of entropy to explain effects on eye movement data[4-6].

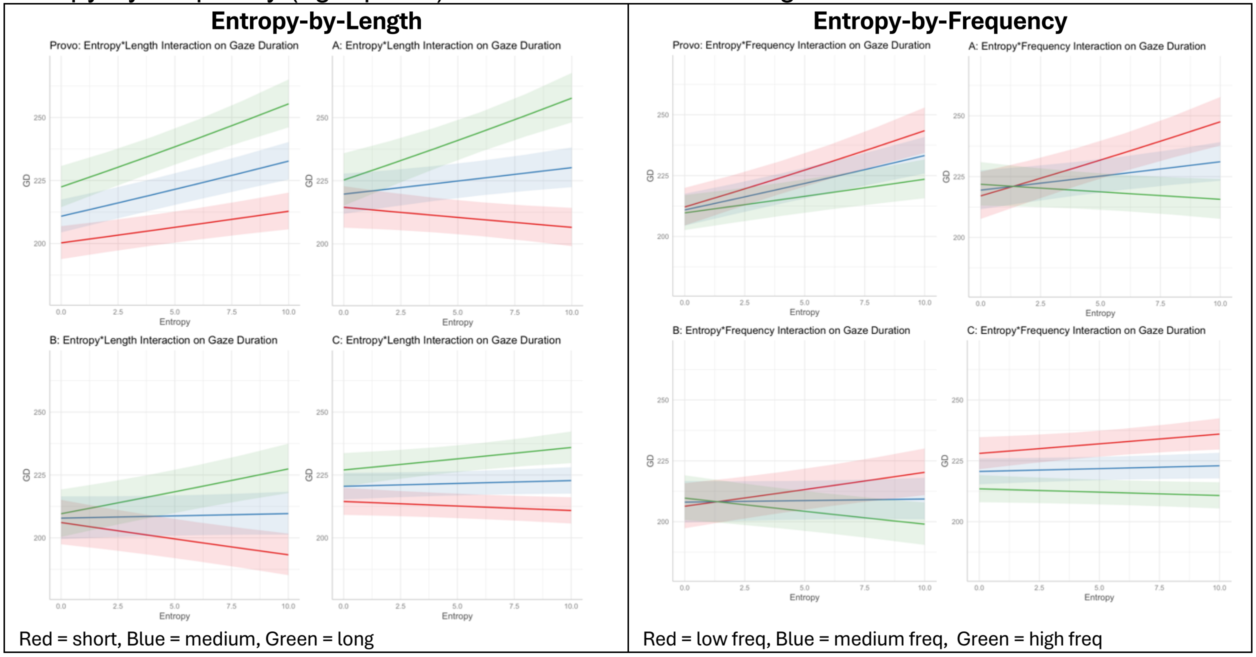
The present study analyzed four datasets containing eye movement data from participants while reading sentences or short texts in English. Three datasets were from previous studies in the authors' lab (A with 90 participants reading 162 sentences, B with 42 participants reading 195 sentences, and C with 181 participants reading 188 sentences); the fourth dataset was the Provo corpus[7] with 84 participants reading 55 short paragraphs. Word-by-word probability distributions were extracted from an LLM (Generative Pre-trained Transformer 2, GPT-2[8]), and these probabilities were used in the calculation of information-theoretic entropy and surprisal. Gaze durations (i.e., the sum of all first-pass fixations) on every word were analyzed as the dependent variable in linear mixed effects models in R version 4.2.2[9] (lme4 package[10]). First, a baseline model was constructed including the fixed effects of length, frequency (SUBTLEX log word frequency[11]), word position in the sentence, word class (i.e., open-class and closed-class), surprisal, and entropy. Random intercepts were included for participant, item, and word, and random slopes for each fixed effect over participant were included in an initial model; if this model did not converge, random slopes explaining minimal variance were removed. Gaze duration was log-transformed, and continuous predictors were centered and scaled. Baseline models (m0) were compared to two additional models: the first model (m1) included an interaction between entropy and length, and the other model (m2) included an interaction between entropy and frequency. Likelihood ratio tests (LRT) were used to assess whether the addition of the interaction significantly improved model fit.

Across all four datasets, both interactions significantly improved model fit and had lower Akaike Information Criterion (AIC) values; see Table 1. The word length effect was very small on constraining contexts (low entropy) but became increasingly larger for more neutral contexts (high entropy); see Figure 1 (left panel). A similar pattern was observed for word frequency, with the magnitude of the effect increasing with entropy; see Figure 1 (right panel). As length and frequency are highly correlated, AIC values for the models with interactions (m1, m2) were also compared. Models with the entropy-by-length interaction had lower AIC values than models with the entropy-by-frequency interaction, although these differences were generally very small. The results highlight the idea that the effects of static characteristics of a word are not independent of the dynamic contexts in which the word can occur. Specifically, the data suggest that readers are particularly sensitive to the effects of length and frequency when the context provides little information about the identity of the upcoming word (i.e., when entropy is high), but that these effects are minimal when the context is highly constraining.

Table 1. LRT and AIC for m0 (baseline) vs. m1 (baseline + entropy*length) & m0 (baseline) vs. m2 (baseline + entropy*frequency). Smaller AIC values for each comparison are italicized.

m0 vs. m1	AIC m0	AIC m1	Chi-square	df	p-value
Provo	93693	*93686*	9.39	1	p < .01
A	57040	*57012*	30.65	1	p < .001
B	30797	*30767*	32.29	1	p < .001
C	226741	*226734*	9.14	1	p < .01
m0 vs. m2	**AIC m0**	**AIC m2**	**Chi-square**	**df**	**p-value**
Provo	93693	*93687*	8.37	1	p < .01
A	57040	*57013*	29.66	1	p < .001
B	30797	*30777*	21.76	1	p < .001
C	226741	*226736*	7.35	1	p < .01

Figure 1. Interaction plots of predicted gaze duration for the entropy-by-length (left panel) and entropy-by-frequency (right panel) interactions. Error shading reflects 95% confidence intervals.



References

1, Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psych Bull*, *124*(3), 372–422.

2. Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6), 641–655.

3. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

4. Cevoli, B., Watkins, C., & Rastle, K. (2022). Prediction as a basis for skilled reading: Insights from modern language models. *Royal Society Open Science*, *9*(6), 211837.

5. Pimentel, T., Meister, C., Wilcox, E. G., Levy, R. P., & Cotterell, R. (2023). On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, *11*, 1624–1642.

6. Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470.

7. Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *BRM*, *50*(2), 826–833.

8. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

9, R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

10. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4* (arXiv:1406.5823). arXix.

11. Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *BRM*, *41*(4), 977–990.