How Well Does First-Token Entropy Approximate Word Entropy as a Psycholinguistic Predictor?

Christian Clark
The Ohio State University
clark.3664@osu.edu

Byung-Doh Oh
New York University
oh.b@nyu.edu

William Schuler
The Ohio State University
schuler.77@osu.edu

Abstract

Contextual entropy is a psycholinguistic measure capturing the anticipated difficulty of processing a word just before it is encountered. Recent studies have tested for entropy-related effects as a potential complement to well-known effects from surprisal. For convenience, entropy is typically estimated based on a language model's probability distribution over a word's first subword token. However, this approximation results in underestimation and potential distortion of true word entropy. To address this, we generate Monte Carlo (MC) estimates of word entropy that allow words to span a variable number of tokens. Regression experiments on reading times show divergent results between first-token and MC word entropy, suggesting a need for caution in using first-token approximations of contextual entropy.

1 Introduction

Recent studies of human sentence processing have explored potential psycholinguistic effects from the contextual entropy of the current word being processed (Cevoli et al., 2022; Pimentel et al., 2023; Wilcox et al., 2023; Giulianelli et al., 2024). Contextual entropy is an information-theoretic measure quantifying a reader's level of uncertainty about the next word based on the current context; it is computed purely based on the conditional probability distribution of the next word without reference to the word's identity. As such, it contrasts with surprisal, another commonly used psycholinguistic measure whose effects can be understood as integration costs for an already observed word (Cevoli et al., 2022; Pimentel et al., 2023).

Contextual entropy is commonly estimated using a language model (LM) like GPT2 (Radford et al., 2019). However, because words can span multiple subword tokens in an LM's vocabulary—making a full summation over their probability distribution intractable—entropy is typically calculated over the probability distribution of each word's first token instead. This practice results in a systematic underprediction of true word entropy (Pimentel et al., 2023), which is magnified in contexts in which multi-token words are probable. This mismatch between the phenomenon of interest (word-level predictive processing) and the metric of choice (token-level entropy) may lead to challenges in drawing psycholinguistic conclusions from experimental results.

To address this issue, we calculate LM-based entropy estimates using a technique based on Monte Carlo (MC; Metropolis and Ulam, 1949) sampling that allows words to span multiple tokens. While this method still only provides an approximation of true word entropy, the MC method can produce unbiased estimates (Giulianelli et al., 2024) whose variance decreases as the number of samples is increased.

To test the difference between first-token entropy and MC word entropy, both measures are evaluated in a set of regression experiments using naturalistic self-paced reading and eye-tracking data. Results show contrasting predictions from the two measures, with MC word entropy making stronger estimates on a corpus of self-paced reading times, but yielding more mixed results on eye-tracking corpora. The gap between the two entropy estimates suggests that first-token entropy may not provide a reliable approximation of true word entropy for psycholinguistic modeling.[1]

2 Related Work

Much of the interest in information-theoretic predictors of human sentence processing traces back to surprisal theory (Hale, 2001; Levy, 2008), which posits a direct link between processing difficulty and a word's surprisal (negative log probability).

[1] Code for first-token and Monte Carlo entropy estimation is available at https://github.com/christian-clark/word-entropy.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 47–57

Robust effects from surprisal have been observed across multiple languages and psycholinguistic measures (e.g., Wilcox et al., 2023; Shain et al., 2024). Entropy-based predictors are thus often evaluated as possible complements to well-established surprisal predictors; we follow this practice by including surprisal predictors in our regression models.

Several forms of entropy have been evaluated in earlier work. Word-level (or token-level) contextual entropy, the predictor of focus in the present study, is studied by van Schijndel and Linzen (2019) as well as several more recent studies (Cevoli et al., 2022; Pimentel et al., 2023; Wilcox et al., 2023; Giulianelli et al., 2024). Other work considers the total entropy of the remainder of a sentence—its raw value (Roark et al., 2009), the reduction in this entropy at each incoming word (Hale, 2003, 2006), or both (Linzen and Jaeger, 2016). Because total entropy is difficult to exactly compute—especially with contemporary language models—some additional work calculates the entropy of a bounded number of future words as a middle ground (Frank, 2013; Giulianelli et al., 2024).

The work by Giulianelli et al. (2024) comes closest to our proposed method of MC estimation of contextual word entropy. These authors approximate the entropy of a continuation of a sentence by generating samples of up to $L \in \{5, 10, 15\}$ tokens following a context string. This differs from our work in that they aim to approximate continuation entropy rather than next-word entropy.

3 Formulations of Contextual Entropy

3.1 Shannon Entropy

Shannon entropy (Shannon, 1948) is the standard form of entropy studied in previous psycholinguistic work. A word's contextual Shannon entropy is defined as its expected surprisal, i.e., its expected negative log probability given the preceding words $w_{1..i-1}$:

$$\mathrm{H}(W_i \mid w_{1..i-1}) = \tag{1}$$
$$-\sum_{w \in V} \mathsf{P}(w \mid w_{1..i-1}) \log_2 \mathsf{P}(w \mid w_{1..i-1}),$$

where V is the vocabulary of all possible words and W_i is a random variable over V.

3.2 Rényi Entropy

Pimentel et al. (2023) discuss the possibility that readers' anticipatory processing may be guided by strategies other than considering the expected surprisal of the next word (i.e., Shannon entropy). For instance, readers might rely on the surprisal of the single most likely word; or, at the other extreme, they might consider the number of possible next words regardless of each word's exact probability.

Rényi entropy (Rényi, 1961) is a generalization of Shannon entropy which captures this spectrum of possible anticipatory reading strategies. A word's contextual Rényi entropy of order α is defined as follows:

$$\mathrm{H}_\alpha(W_i \mid w_{1..i-1}) = \tag{2}$$
$$\lim_{\beta \to \alpha} \frac{1}{1-\beta} \log_2 \sum_{w \in V} \Big(\mathsf{P}(w \mid w_{1..i-1}) \Big)^\beta.$$

When $\alpha = 0$, Rényi entropy considers the number of possible next words. If $\alpha = 1$, Rényi entropy equals Shannon entropy (via an application of L'Hôpital's rule to Eq. (2)). And when $\alpha = \infty$, Rényi entropy measures the surprisal of the single most likely word in context.

Pimentel et al. (2023) test several values of α and find that Rényi contextual entropy with $\alpha = 1/2$ is a relatively strong predictor of reading times. Our regression experiments evaluate both Shannon entropy and Rényi entropy with $\alpha = 1/2$ to get a fuller picture of how first-token and MC approximation affect metrics describing anticipatory word processing.

3.3 First-Token and MC Approximations

Contemporary LMs typically work with a finite subword vocabulary—defined using a method like Byte-Pair Encoding (Sennrich et al., 2016)—that supports an infinite word vocabulary, as words can span a variable number of subwords. Because summing over an infinite vocabulary is intractable, recent work instead considers the entropy of a word's first subword token. First-token contextual Shannon entropy is defined as follows:

$$\mathrm{H}(W_i \mid w_{1..i-1}) \approx \tag{3}$$
$$-\sum_{t \in T} \mathsf{P}(t \mid w_{1..i-1}) \log_2 \mathsf{P}(t \mid w_{1..i-1}),$$

where t is a token drawn from the LM's subword vocabulary T. First-token Rényi entropy can be defined analogously. These approximation are a

lower bound on true word entropy (Pimentel et al., 2023).

MC estimation of contextual word entropy is performed by computing the average surprisal of a multiset of words S sampled in context $w_{1..i-1}$:

$$\mathrm{H}(W_i \mid w_{1..i-1}) \approx -\frac{1}{|S|} \sum_{s \in S} \log_2 \mathsf{P}(s \mid w_{1..i-1}). \tag{4}$$

Each sampled word $s \in S$ is produced by randomly generating successive subword tokens until a word boundary (i.e., the subsequent whitespace in English) is reached. The surprisal of s is calculated by summing over the surprisal of its subword tokens. This form of MC estimation is unbiased because it directly averages over samples drawn from the ground-truth probability distribution $\mathsf{P}(s \mid w_{1..i-1})$ (Giulianelli et al., 2024).

An MC estimate of Rényi entropy can be obtained by replacing the summation term in Equation (2) with a sample-based approximation:

$$\mathrm{H}_\alpha(W_i \mid w_{1..i-1}) \approx \tag{5}$$
$$\lim_{\beta \to \alpha} \frac{1}{1-\beta} \log_2 \left(\frac{1}{|S|} \sum_{s \in S} \left(\mathsf{P}(s \mid w_{1..i-1}) \right)^{\beta-1} \right).$$

Because of the log-transformation of the sample-based estimate, Equation (5) is not an unbiased estimate of true Rényi entropy.[2] However, in a randomized trial we found that this approximation of Rényi entropy results in negligible bias when $|S| \geq 64$.

Due to limitations in the available compute budget, all MC estimates in this work restrict the number of samples to $|S| = 512$. An analysis of sample variance indicated that entropy estimates are reasonably stable with this sample count. See Appendix A for this analysis and other details about the MC estimation.

To illustrate the difference between first-token and MC word entropy, Figure 1 shows the average Shannon entropies of the 10 most frequent parts of speech in the Natural Stories corpus (Futrell et al., 2021), one of the English psycholinguistic corpora used in subsequent regression experiments (Sec. 4). As expected, MC word entropy is consistently higher than first-token entropy. It can also be observed that open-class parts of speech such as NN (noun) and JJ (adjective) show a larger relative difference between first-token and MC word entropy compared to closed classes like IN (preposition). This likely reflects a wider range of multi-token words available within these part-of-speech categories; it also provides evidence that the first-token approximation not only underpredicts word entropy but also distorts the difference between word classes.

[2]This is due to Jensen's inequality, which shows that in general, $f(\mathbb{E}(x)) \neq \mathbb{E}(f(x))$, where $\mathbb{E}()$ is an expectation and f in this case is $\log_2$.

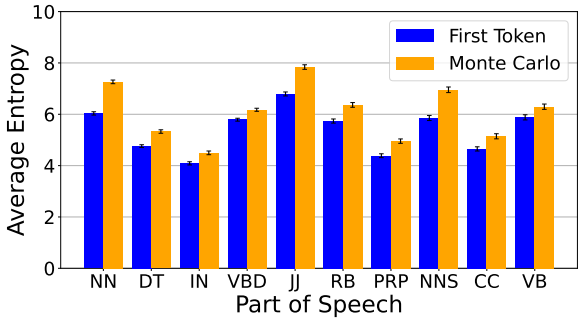


Figure 1: Average Shannon entropy of the 10 most frequent parts of speech in the Natural Stories corpus, using either first-token or Monte Carlo word entropy approximations. Part-of-speech tags are from Penn Treebank annotations (Marcus et al., 1993). Error bars represent ±1 standard error of the mean (SEM).

Corpus	Observations
Natural Stories SPR	770,259
Brown SPR	119,120
Dundee FP	195,507
Dundee GP	195,507
Provo FP	106,138
Provo GP	106,139
GECO FP	289,892
GECO GP	289,892

Table 1: The number of observations in each reading-time corpus. These observations were partitioned into 10 folds of roughly equal sizes for cross-validation.

4 Regression Experiments

To compare the psycholinguistic predictive power of first-token entropy and word entropy, we perform linear mixed-effects (LME; Bates et al., 2015) regression experiments on a set of naturalistic reading-time corpora in English.

4.1 Corpora

The psycholinguistic corpora included two self-paced reading (SPR) corpora and three corpora with first-pass (FP) and go-past (GP) durations from eye tracking. The self-paced reading corpora were Natural Stories (Futrell et al., 2021), which

contains data from 181 subjects who read 10 naturalistic stories; and Brown (Smith and Levy, 2013), which contains data from 35 subjects who read 13 passages from the Brown corpus. The eye-tracking corpora were Dundee (Kennedy et al., 2003), containing fixation durations from 10 subjects who read 67 newspaper editorials; Provo (Luke and Christianson, 2018), containing fixation durations from 84 subjects who read 55 passages from a variety of sources including news articles, magazines, and works of fiction; and GECO (Cop et al., 2017), containing fixation durations from 14 subjects who read a 13-chapter Agatha Christie novel (Christie, 1920). Table 1 shows the number of observations per partition in each corpus.

4.2 Predictors

Baseline predictors in the LME models included word length, word index, unigram surprisal, LM surprisal of the current and previous word (SPR, FP, and GP), and whether the previous word was fixated (FP and GP only). Unigram surprisal was estimated on the 6.5 billion whitespace-delimited words from the OpenWebText Corpus (Gokaslan and Cohen, 2019) using the KenLM toolkit (Heafield et al., 2013). SPR regression models included per-subject random slopes for word length, word index, and LM surprisal (current and previous word), and a per-subject random intercept. Regression models for eye-tracking only included a per-subject random intercept. Other random slopes were removed to ensure convergence.

4.3 Evaluation

For each corpus and response type, k-fold cross-validation with $k = 10$ was used to find an average difference in log likelihood (ΔLL) between a regression model containing only the baseline predictors, and a regression model additionally containing an entropy predictor (either first-token entropy or MC word entropy).[3] This evaluation was conducted twice, once using Shannon entropy and once using Rényi entropy ($\alpha = 1/2$). GPT2-small was the LM used to calculate entropy and surprisal predictors. The full ΔLL for each fold was divided by its number of datapoints to calculate a per-datapoint ΔLL value.

Paired permutation tests over the ΔLL values from each fold were conducted to determine whether the differences between models using first-token and MC word entropy were statistically significant. Aggregated significance tests for Shannon and Rényi entropy were also performed using a paired permutation test over the concatenated ΔLL values across the 10 folds of each corpus.

[3] Each observation was assigned to partition $p = (m + n) \bmod 10$, where m and n are subject ID and sentence number respectively.

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$
NS SPR	1.26e−4	3.05e−4 ***
Brown SPR	2.10e−6	−1.75e−5
Dundee FP	1.28e−5	−4.09e−6
Dundee GP	9.21e−6	−4.60e−6
Provo FP	−7.54e−6	1.18e−5
Provo GP	2.13e−4 **	6.48e−5
GECO FP	8.11e−5 ***	1.17e−5
GECO GP	−1.38e−6	−3.45e−6
Combined	7.09e−5	1.17e−4 *

(a) Shannon Entropy

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$
NS SPR	2.61e−4	1.43e−3 ***
Brown SPR	−1.05e−5	1.99e−4
Dundee FP	1.02e−6	1.10e−4 **
Dundee GP	−4.09e−6	1.90e−4 **
Provo FP	−1.29e−5	−8.95e−6
Provo GP	4.51e−5	6.57e−5
GECO FP	5.86e−6	2.76e−5
GECO GP	0.00	1.14e−5
Combined	9.85e−5	5.78e−4 ***

(b) Rényi Entropy

Table 2: Increases in per-datapoint log likelihood (measured in nats) from adding a target entropy predictor to a baseline regression model for predicting self-paced reading (SPR) time, first-pass (FP) duration, or go-past (GP) duration. $\Delta\text{LL}_{\text{FT}}$ and $\Delta\text{LL}_{\text{MC}}$ respectively refer to log likelihood improvements from first-token and MC entropy approximations. NS means Natural Stories. Sigificance levels are * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.4 Results

Table 2 presents the results from this experiment. When using Shannon entropy (Table 2a), replacing first-token estimates with MC estimates improves ΔLL scores on the Natural Stories self-paced reading corpus. However, results are less consistent across other corpora, with two cases in which first-token entropy significantly outperforms MC word entropy. Nonetheless, the improvement from MC word entropy on Natural Stories is strong enough to lead to a significant improvement in the aggregated permutation test.

The results using Rényi entropy (Table 2b) show

a more robust improvement in reading predictions from MC word entropy. Higher ΔLL values are observed across all corpora, with significant improvements from MC word entropy on several corpora and in the aggregated evaluation. These results suggest that—unlike Shannon entropy—word-level estimates of Rényi entropy are consistently better predictors of anticipatory processing.

Appendix B presents the effect sizes of the entropy predictors as well as additional regression results. In general, it can be seen that entropy predictors have smaller effects than surprisal predictors (B.1, B.5), and that removing surprisal predictors generally increases the improvements from entropy predictors (B.4). It can also be observed that mean squared error and R^2, two alternative measures of prediction quality, show generally consistent trends to those reported for ΔLL (B.2), and that results are similar when entropy is calculated using the OPT-125M (Zhang et al., 2022) or Pythia-160M (Biderman et al., 2023) LMs instead of GPT2-small (B.3).

5 Conclusion

First-token approximations of contextual entropy have often been used as a tool to study anticipatory word-level processing (Cevoli et al., 2022; Pimentel et al., 2023; Wilcox et al., 2023). However, this work shows that estimates of word entropy from MC sampling often lead to experimental results that diverge from those using first-token entropy; in many cases, MC estimates provide a closer match to human behavioral data. The concrete difference across the two conditions warrants caution against using first-token entropy in psycholinguistic modeling.

Limitations

Concerns about first-token entropy estimates are relevant for contemporary Transformer LMs that use a vocabulary of subword tokens; however, they do not apply to word-level LMs, such as the LSTM used by van Schijndel and Linzen (2019) to study contextual entropy.

This paper considers entropy measures for English based on GPT2-small. However, the degree to which first-token entropy distorts true word entropy may vary depending the language of study and LM of choice (although we saw reasonably stable results across two alternative English LMs, as reported in Appendix B.3). For instance, LMs with larger subword vocabularies may have a closer to 1:1 relationship between tokens and words and thus show less distortion. It is also possible that languages with higher morphological complexity than English (e.g., agglutinative languages) will tend to have more multi-token words and therefore will show a larger gap between first-token and word entropy.

One further limitation is the higher cost of calculating MC estimates of word entropy relative to first-token entropy. Computing MC entropy estimates based on $|S| = 512$ word samples took approximately 30 seconds per sentence and required roughly 350 GB of memory.[4] While this cost was manageable for the medium-scale psycholinguistic corpora in this study, calculating MC entropy estimates for larger-scale corpora used in other natural language processing applications may be less practical.

Acknowledgments

We thank the ARR reviewers and the area chair for their helpful comments. This work was supported by the National Science Foundation (NSF) grant #2313140. All views expressed are those of the authors and do not necessarily reflect the views of the NSF. Computations for this work were partly run using the Ohio Supercomputer Center (1987). This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 2397–2430.

Benedetta Cevoli, Chris Watkins, and Kathleen Rastle. 2022. Prediction as a basis for skilled reading: Insights from modern language models. *Royal Society open science*, 9(6):211837.

[4] We computed batches of 8 next-word samples concurrently; this batching reduces runtime but increases memory demand.

Agatha Christie. 1920. *The mysterious affair at Styles*. John Lane. Retrieved from Project Gutenberg.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

Stefan L Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in cognitive science*, 5(3):475–494.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11648–11669, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics*, pages 159–166, Pittsburgh, PA.

John Hale. 2003. The information conveyed by words in sentences. *Journal of psycholinguistic research*, 32(2):101–123.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive science*, 30(4):643–672.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696.

Alan Kennedy, James Pynte, and Robin Hill. 2003. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen and T Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive science*, 40(6):1382–1411.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Nicholas Metropolis and Stanislaw Ulam. 1949. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

Ohio Supercomputer Center. 1987. Ohio Supercomputer Center.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Ethan G Wilcox, Roger P Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *ArXiv*.

Alfréd Rényi. 1961. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 324–333.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Marten van Schijndel and Tal Linzen. 2019. Can entropy explain successor surprisal effects in reading? In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 1–7.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

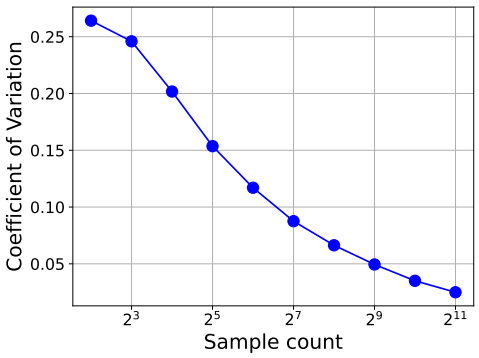
A Monte Carlo Sampling Details

A.1 Sampling Procedure

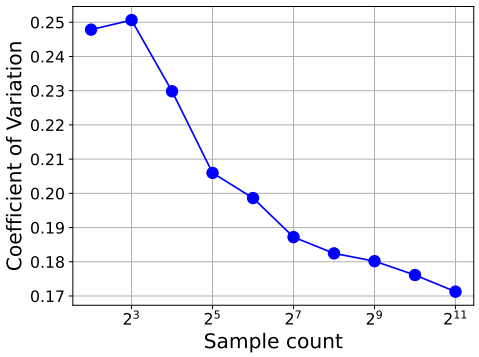
Sampling a potentially multi-token word w_i from a language model's conditional probability distribution $\mathsf{P}(W_i \mid w_{1..i-1})$ involves iteratively sampling the subword tokens of w_i until a word boundary is reached. A challenge that arises is that LMs like GPT2 treat word boundaries as leading whitespaces on tokens, meaning that the end-of-word boundary for w_i will be part of the first token for w_{i+1}. These word boundaries must be carefully managed in order to ensure that probabilities of multi-token words form a proper distribution (Oh and Schuler, 2024; Pimentel and Meister, 2024).

To properly track word boundaries, we follow Oh and Schuler (2024) in separating the subword vocabulary T of an LM into a subset containing whitespace-initial tokens T_B and a subset containing tokens with no initial whitespace T_I. Note that $T = T_B \cup T_I$ and $T_B \cap T_I = \emptyset$. To sample w_i, we first sample a word-initial token from T_B, with LM probabilities renormalized to sum to 1 over T_B. Subsequent tokens are drawn from $T_I \cup \{\text{EOW}\}$, where tokens in T_I are assigned their usual conditional probabilities, and

$$\mathsf{P}(\text{EOW} \mid w_{i..i-1}, t_{1..j}) = \sum_{t \in T_B} \mathsf{P}(t \mid w_{i..i-1}, t_{1..j}),$$



(a) Shannon Entropy



(b) Rényi Entropy

Figure 2: Coefficient of variation by sample count for Monte Carlo estimates of word entropy.

where $t_{1..j}$ are the subword tokens that have been sampled so far. In other words, the probability of EOW is the probability of sampling any whitespace-initial token, which means the end of w_i has been reached. The surprisal of w_i is the sum of the surprisal of the initial token from T_B, any middle tokens from T_I, and the final EOW.

To ensure that sampling was tractable in our experiments, the number of possible subword tokens in a word was capped at 20. Strictly speaking, this means that the MC estimates in this work will tend to underestimate true surprisal. However, words with more than 20 tokens are exceptionally unlikely to sample—empirically, we observe that around 1/50,000 samples reach 20 tokens—so the effect on MC estimates should be minimal.

A.2 Variance Analysis

To measure the effect of sample count on the variance of the Monte Carlo estimates, we performed a bootstrapping (Efron, 1992) analysis using the first story from the Natural Stories corpus. We followed a similar procedure to Giulianelli et al. (2024, Sec. 5.1). First, for each $k \in \{2^j \mid j =$

Corpus	Shannon Entropy FT	Shannon Entropy MC	Rényi Entropy FT	Rényi Entropy MC
NS SPR	1.9752	3.7466	3.0044	6.0325
Brown SPR	−0.0583	0.6345	−0.2484	1.4915
Dundee FP	−0.8926	−0.3298	−0.5853	2.0721
Dundee GP	−1.5943	−0.2209	−0.5138	5.1233
Provo FP	−1.0993	−1.6591	0.2369	0.0957
Provo GP	−9.8869	−5.7522	−4.898	4.8992
GECO FP	1.5739	0.6698	0.4851	−0.9177
GECO GP	−0.4139	0.1021	0.5646	1.7777

Table 3: Effect sizes (ms) from regression predictors based on first-token (FT) and Monte Carlo (MC) approximations of word entropy.

$2, 3, \ldots, 11\}$ and word position i in the story, a set of k word samples $S_{k,i}$ was taken following the procedure in A.1. Next, a set of 1000 resamples with replacement—each of size k—were taken from each $S_{k,i}$, and entropy was calculated over each resample. The coefficient of variation for word i was then calculated as $CV_{k,i} = \sigma_{k,i}/\mu_{k,i}$, where $\sigma_{k,i}$ and $\mu_{k,i}$ are respectively the standard deviation and mean of the entropy across all resamples. Finally, the average coefficient of variation with k samples, CV_k, was found by averaging over all $CV_{k,i}$ values.

Figure 2 plots the CV_k value for each sample size k for both Shannon and Rényi entropy. Generally speaking, Rényi entropy requires more samples than Shannon entropy to attain a given coefficient of variation. Both entropy measures benefit from higher sample counts, but the CV_k values are reasonably stable near the sample count of $2^9 = 512$ used in the regression experiments (Sec. 4).

B Additional Regression Results

B.1 Effect Sizes from GPT2 Entropy

Table 3 shows the effect sizes from the GPT2-based entropy predictors evaluated in Section 4. The effect sizes are averaged over the 10-fold cross-validation models.

B.2 MSE and R^2 from GPT2 Entropy

Table 4 presents several measures from the regression models tested in main experiments. Along with the ΔLL measures discussed in Section 4.2, this table includes the following measures:

- MSE_{FT}, MSE_{MC}: mean squared error from regression models containing first-token and Monte Carlo estimates of entropy
- $\Delta\text{MSE}_{\text{FT}}$, $\Delta\text{MSE}_{\text{MC}}$: changes in mean squared error relative to a baseline regression model with no entropy predictor
- R^2_{FT}, R^2_{MC}: coefficients of determination from regression models containing first-token and Monte Carlo estimates of entropy
- ΔR^2_{FT}, ΔR^2_{MC}: changes in R^2 relative to a baseline regression model with no entropy predictor

Like ΔLL, each of these measures is averaged over results from 10-fold cross-validation.

B.3 Additional Language Models

To check whether the results in Section 4.4 generalize beyond the GPT2-small language model, we collected regression results using entropy measures from two additional language models. Table 5 presents results from the OPT-125M language model (Zhang et al., 2022), and Table 6 presents results from the Pythia-160M language model (Biderman et al., 2023).

B.4 No Surprisal Control

Table 7 presents results from regression models that use entropy estimates from GPT2-small but no surprisal control. Comparing these results with Table 4 shows the degree to which including a surprisal control weakens effects from entropy.

B.5 Surprisal-Only Regression Model

Table 8 shows the effect sizes and improvements in regression model fit due to GPT2 surprisal, in order to give a frame of reference for improvements from entropy. LL_{surp}, MSE_{surp}, and R^2_{surp} values come from a regression model with GPT2 surprisal predictor as well as the other baseline predictors listed in Section 4.2. $\Delta\text{LL}_{\text{surp}}$, $\Delta\text{MSE}_{\text{surp}}$, and ΔR^2_{surp} compare this model to a baseline model with no surprisal predictor. For this evaluation, neither the baseline nor the main effect model included an entropy predictor.

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$	MSE_{FT}	MSE_{MC}	$\Delta\text{MSE}_{\text{FT}}$	$\Delta\text{MSE}_{\text{MC}}$	R^2_{FT}	R^2_{MC}	ΔR^2_{FT}	ΔR^2_{MC}
NS SPR	1.26e−4	3.05e−4	20 077	20 070	−6.0094	−13.4711	0.2708	0.2711	2.18e−4	4.89e−4
Brown SPR	2.10e−6	−1.75e−5	21 893	21 891	−3.1877	−4.3147	0.2238	0.2239	1.13e−4	1.53e−4
Dundee FP	1.28e−5	−4.09e−6	17 012	17 013	−0.4241	0.1208	0.126	0.126	2.18e−5	−6.21e−6
Dundee GP	9.21e−6	−4.60e−6	62 203	62 205	−1.1001	0.6039	0.0576	0.0576	1.67e−5	−9.15e−6
Provo FP	−7.54e−6	1.18e−5	36 175	36 173	0.5538	−0.8149	0.1091	0.1092	−1.36e−5	2.01e−5
Provo GP	2.13e−4	6.48e−5	150 401	150 447	−70.3786	−24.3149	0.0709	0.0706	4.35e−4	1.50e−4
GECO FP	8.11e−5	1.17e−5	11 973	11 975	−1.9361	−0.2672	0.1436	0.1435	1.38e−4	1.91e−5
GECO GP	−1.38e−6	−3.45e−6	91 133	91 133	0.1751	0.5549	0.0564	0.0564	−1.81e−6	−5.75e−6
Combined	7.09e−5	1.17e−4	40 171	40 171	−6.3828	−6.4331	0.1569	0.1569	1.34e−4	1.35e−4

(a) Shannon Entropy

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$	MSE_{FT}	MSE_{MC}	$\Delta\text{MSE}_{\text{FT}}$	$\Delta\text{MSE}_{\text{MC}}$	R^2_{FT}	R^2_{MC}	ΔR^2_{FT}	ΔR^2_{MC}
NS SPR	2.61e−4	1.43e−3	20 069	20 028	−14.5593	−55.574	0.2711	0.2726	5.29e−4	2.02e−3
Brown SPR	−1.05e−5	1.99e−4	21 890	21 883	−5.9747	−12.8295	0.2239	0.2242	2.12e−4	4.55e−4
Dundee FP	1.02e−6	1.10e−4	17 013	17 009	−0.0389	−3.7241	0.126	0.1262	2.00e−6	1.91e−4
Dundee GP	−4.09e−6	1.90e−4	62 205	62 181	0.5899	−23.406	0.0576	0.0579	−8.94e−6	3.55e−4
Provo FP	−1.29e−5	−8.95e−6	36 175	36 175	0.9899	0.6182	0.1091	0.1091	−2.44e−5	−1.52e−5
Provo GP	4.51e−5	6.57e−5	150 454	150 453	−17.3379	−17.6413	0.0705	0.0705	1.07e−4	1.09e−4
GECO FP	5.86e−6	2.76e−5	11 975	11 974	−0.128	−0.6585	0.1435	0.1435	9.16e−6	4.71e−5
GECO GP	0.00	1.14e−5	91 132	91 130	−0.0233	−2.0846	0.0564	0.0565	2.42e−7	2.16e−5
Combined	9.85e−5	5.78e−4	40 171	40 152	−6.561	−25.2072	0.1569	0.1573	1.38e−4	5.29e−4

(b) Rényi Entropy

Table 4: Regression results using first-token (FT) and Monte Carlo (MC) entropy estimates from GPT2-small.

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$	MSE_{FT}	MSE_{MC}	$\Delta\text{MSE}_{\text{FT}}$	$\Delta\text{MSE}_{\text{MC}}$	R^2_{FT}	R^2_{MC}	ΔR^2_{FT}	ΔR^2_{MC}
NS SPR	1.61e−4	4.15e−4	20 076	20 064	−6.0137	−17.403	0.2708	0.2713	2.18e−4	6.32e−4
Brown SPR	−1.05e−4	−4.65e−5	21 896	21 893	−0.2538	−3.0596	0.2237	0.2238	9.00e−6	1.08e−4
Dundee FP	5.11e−6	−3.07e−6	17 009	17 009	−0.1497	0.088	0.1262	0.1262	7.69e−6	−4.52e−6
Dundee GP	6.65e−6	−3.07e−6	62 206	62 207	−0.7871	0.4156	0.0576	0.0575	1.19e−5	−6.30e−6
Provo FP	7.82e−6	6.03e−5	36 176	36 172	−0.4543	−3.9773	0.1091	0.1092	1.12e−5	9.79e−5
Provo GP	2.27e−4	9.76e−5	150 389	150 429	−72.3363	−32.4111	0.0709	0.0707	4.47e−4	2.00e−4
GECO FP	1.54e−4	4.17e−5	11 975	11 978	−3.7058	−1.0149	0.1435	0.1433	2.65e−4	7.26e−5
GECO GP	−1.38e−6	−1.72e−6	91 141	91 141	0.2182	0.2928	0.0564	0.0564	−2.26e−6	−3.03e−6
Combined	8.82e−5	1.65e−4	40 171	40 169	−6.5538	−8.5611	0.1569	0.1569	1.38e−4	1.80e−4

(a) Shannon Entropy

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$	MSE_{FT}	MSE_{MC}	$\Delta\text{MSE}_{\text{FT}}$	$\Delta\text{MSE}_{\text{MC}}$	R^2_{FT}	R^2_{MC}	ΔR^2_{FT}	ΔR^2_{MC}
NS SPR	3.16e−4	1.50e−3	20 067	20 025	−14.7888	−56.4398	0.2712	0.2727	5.37e−4	2.05e−3
Brown SPR	−8.95e−5	5.47e−5	21 894	21 891	−1.4869	−4.7233	0.2238	0.2239	5.27e−5	1.67e−4
Dundee FP	−5.11e−7	7.67e−6	17 009	17 009	0.0282	−0.2357	0.1262	0.1262	−1.45e−6	1.21e−5
Dundee GP	−5.11e−6	2.71e−5	62 207	62 203	0.6228	−3.3687	0.0575	0.0576	−9.44e−6	5.10e−5
Provo FP	−8.01e−6	−1.23e−5	36 177	36 177	0.5607	0.9029	0.1091	0.1091	−1.38e−5	−2.22e−5
Provo GP	6.55e−5	1.93e−5	150 439	150 456	−22.9908	−5.3332	0.0706	0.0705	1.42e−4	3.29e−5
GECO FP	5.00e−5	5.86e−6	11 977	11 978	−1.2196	−0.144	0.1433	0.1432	8.72e−5	1.03e−5
GECO GP	2.41e−6	6.21e−6	91 140	91 140	−0.5322	−1.0788	0.0564	0.0564	5.51e−6	1.12e−5
Combined	1.22e−4	5.64e−4	40 171	40 156	−6.9143	−21.9862	0.1569	0.1572	1.45e−4	4.61e−4

(b) Rényi Entropy

Table 5: Regression results using first-token (FT) and Monte Carlo (MC) entropy estimates from OPT-125M.

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$	MSE_{FT}	MSE_{MC}	$\Delta\text{MSE}_{\text{FT}}$	$\Delta\text{MSE}_{\text{MC}}$	R^2_{FT}	R^2_{MC}	ΔR^2_{FT}	ΔR^2_{MC}
NS SPR	1.44e−4	3.64e−4	20 079	20 069	−6.4502	−16.5472	0.2707	0.2711	2.34e−4	6.01e−4
Brown SPR	−2.28e−5	3.27e−6	21 920	21 918	−0.2588	−2.0246	0.2229	0.2229	9.18e−6	7.18e−5
Dundee FP	1.33e−5	0.00	17 018	17 019	−0.4729	−0.0079	0.1257	0.1257	2.43e−5	4.07e−7
Dundee GP	2.30e−5	1.53e−6	62 228	62 231	−2.8315	−0.2539	0.0572	0.0572	4.29e−5	3.85e−6
Provo FP	−8.20e−6	1.35e−5	36 163	36 161	0.6008	−0.8809	0.1094	0.1095	−1.48e−5	2.17e−5
Provo GP	1.12e−4	2.56e−5	150 498	150 526	−37.6887	−9.4808	0.0703	0.0701	2.33e−4	5.86e−5
GECO FP	8.49e−5	1.66e−5	11 985	11 987	−2.0528	−0.3869	0.1427	0.1426	1.47e−4	2.77e−5
GECO GP	−1.38e−6	6.90e−7	91 172	91 172	0.245	−0.0412	0.056	0.056	−2.54e−6	4.27e−7
Combined	7.25e−5	1.40e−4	40 187	40 185	−4.8762	−6.8816	0.1566	0.1566	1.02e−4	1.44e−4

(a) Shannon Entropy

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$	MSE_{FT}	MSE_{MC}	$\Delta\text{MSE}_{\text{FT}}$	$\Delta\text{MSE}_{\text{MC}}$	R^2_{FT}	R^2_{MC}	ΔR^2_{FT}	ΔR^2_{MC}
NS SPR	3.82e−4	1.92e−3	20 066	20 020	−19.3132	−64.9823	0.2712	0.2729	7.01e−4	2.36e−3
Brown SPR	−5.14e−5	1.75e−4	21 920	21 908	0.6322	−11.878	0.2228	0.2233	−2.24e−5	4.21e−4
Dundee FP	0.00	−1.53e−6	17 019	17 019	−0.0135	0.0349	0.1257	0.1257	6.92e−7	−1.79e−6
Dundee GP	1.49e−16	0.00	62 231	62 231	−0.0147	−0.1743	0.0572	0.0572	2.22e−7	2.64e−6
Provo FP	−2.64e−6	1.72e−5	36 162	36 162	0.2361	−0.7079	0.1094	0.1095	−5.81e−6	1.74e−5
Provo GP	1.54e−5	3.77e−7	150 529	150 535	−6.4558	−0.518	0.0701	0.07	3.99e−5	3.20e−6
GECO FP	3.38e−5	1.66e−5	11 986	11 987	−0.8126	−0.4139	0.1427	0.1426	5.81e−5	2.96e−5
GECO GP	3.79e−6	2.38e−5	91 171	91 168	−0.6358	−4.0685	0.056	0.0561	6.58e−6	4.21e−5
Combined	1.45e−4	7.30e−4	40 185	40 167	−7.6655	−25.5373	0.1566	0.157	1.61e−4	5.36e−4

(b) Rényi Entropy

Table 6: Regression results using first-token (FT) and Monte Carlo (MC) entropy estimates from Pythia-160M.

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$	MSE_{FT}	MSE_{MC}	$\Delta\text{MSE}_{\text{FT}}$	$\Delta\text{MSE}_{\text{MC}}$	R^2_{FT}	R^2_{MC}	ΔR^2_{FT}	ΔR^2_{MC}
NS SPR	2.47e−4	5.40e−4	20 337	20 327	−12.2938	−22.4254	0.2613	0.2617	4.47e−4	8.14e−4
Brown SPR	5.17e−5	1.94e−4	22 543	22 536	−6.7647	−14.0904	0.2008	0.201	2.40e−4	5.00e−4
Dundee FP	3.32e−4	4.18e−4	17 189	17 186	−11.4406	−14.3736	0.1169	0.1171	5.88e−4	7.38e−4
Dundee GP	1.29e−4	2.11e−4	62 530	62 519	−16.3992	−26.6459	0.0526	0.0528	2.48e−4	4.04e−4
Provo FP	2.09e−4	1.35e−4	36 420	36 426	−17.046	−11.7245	0.1031	0.103	4.20e−4	2.89e−4
Provo GP	1.88e−7	4.56e−5	151 191	151 178	0.0633	−12.7081	0.066	0.0661	−3.91e−7	7.85e−5
GECO FP	9.35e−4	5.71e−4	12 082	12 091	−22.6558	−13.8392	0.1358	0.1352	1.62e−3	9.90e−4
GECO GP	1.73e−4	1.81e−4	91 596	91 595	−31.9843	−33.1864	0.0516	0.0517	3.31e−4	3.44e−4
Combined	3.04e−4	3.86e−4	40 485	40 481	−16.097	−20.8434	0.1503	0.1504	3.38e−4	4.37e−4

(a) Shannon Entropy

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$	MSE_{FT}	MSE_{MC}	$\Delta\text{MSE}_{\text{FT}}$	$\Delta\text{MSE}_{\text{MC}}$	R^2_{FT}	R^2_{MC}	ΔR^2_{FT}	ΔR^2_{MC}
NS SPR	4.90e−4	1.91e−3	20 325	20 278	−24.1903	−71.3104	0.2618	0.2635	8.79e−4	2.59e−3
Brown SPR	1.40e−4	5.92e−4	22 535	22 520	−14.5117	−29.5981	0.201	0.2016	5.15e−4	1.05e−3
Dundee FP	4.48e−4	8.26e−4	17 185	17 172	−15.4315	−28.3935	0.1171	0.1178	7.93e−4	1.46e−3
Dundee GP	2.28e−4	6.89e−4	62 517	62 460	−28.7317	−86.4387	0.0528	0.0537	4.35e−4	1.31e−3
Provo FP	−1.39e−2	1.89e−4	36 713	36 420	275.5705	−17.0949	0.0959	0.1031	−6.79e−3	4.21e−4
Provo GP	9.28e−5	3.99e−4	151 164	151 068	−26.3901	−122.6861	0.0661	0.0667	1.63e−4	7.58e−4
GECO FP	5.48e−4	6.28e−5	12 091	12 103	−13.3075	−1.5096	0.1352	0.1343	9.52e−4	1.08e−4
GECO GP	2.18e−4	1.77e−4	91 588	91 596	−40.1815	−32.2722	0.0517	0.0516	4.16e−4	3.34e−4
Combined	−3.46e−4	9.50e−4	40 493	40 451	−8.7115	−50.9217	0.1502	0.151	1.83e−4	1.07e−3

(b) Rényi Entropy

Table 7: Results from a regression models without a surprisal control. First-token (FT) and Monte Carlo (MC) entropy estimates are from GPT2-small.

Corpus	Effect Size	$\Delta\text{LL}_{\text{surp}}$	MSE_{surp}	$\Delta\text{MSE}_{\text{surp}}$	R^2_{surp}	ΔR^2_{surp}
NS SPR	5.0588	8.06e−4	20 320	−29.1578	0.262	1.06e−3
Brown SPR	7.4358	1.07e−3	22 510	−39.7916	0.2019	1.41e−3
Dundee FP	12.811	4.07e−3	17 063	−137.5839	0.1234	7.07e−3
Dundee GP	16.9272	1.91e−3	62 284	−261.8856	0.0564	3.97e−3
Provo FP	14.9367	5.38e−3	36 105	−332.7122	0.1109	8.19e−3
Provo GP	27.0395	2.49e−3	150 431	−759.4566	0.0707	4.69e−3
GECO FP	12.1977	5.86e−3	11 960	−144.4364	0.1445	1.03e−2
GECO GP	21.3154	1.77e−3	91 283	−344.5515	0.0549	3.57e−3
Combined	—	2.39e−3	40 326	−175.1417	0.1536	3.68e−3

Table 8: Effect size (ms) and improvements in regression model fit from GPT-2 surprisal, with no entropy predictors included.