Improved Reading Time Predictions from Word-Level Contextual Entropy

Christian Clark,[1] Byung-Doh Oh,[2] William Schuler[1]

[1]Ohio State University; [2]New York University

clark.3664@osu.edu

Background. Recent studies of human sentence processing have reported evidence of psycholinguistic effects from the contextual entropy $\mathrm{H}(W_i \mid w_{1..i-1})$ of the current word W_i being processed [1, 2, 3]. A word's contextual entropy is defined as its expected surprisal: $\mathrm{H}(W_i \mid w_{1..i-1}) = -\sum_{w \in V} \mathrm{P}(w \mid w_{1..i-1}) \log_2 \mathrm{P}(w \mid w_{1..i-1})$, where V is the vocabulary. This measure captures the predictive processing difficulty before each new word is encountered, in contrast to raw surprisal, whose effects can be understood as integration costs for an already observed word [1, 4]. Previous work estimates contextual entropy using a language model (LM) like GPT2 [5]. However, because words can span multiple subword tokens in an LM's vocabulary—and therefore are intractable to sum probabilities over—entropy is typically calculated over each word's first token instead. This practice results in a systematic underprediction of true word entropy [1], which is magnified in contexts in which multi-token words are probable. To address this issue, we calculate LM-based entropy estimates using a Monte Carlo (MC; [6]) technique that randomly samples token sequences to explore, and thus allows words to span multiple tokens. We then evaluate the fit of the MC estimates to naturalistic reading times.

Methods. Mixed-effects regression experiments were conducted on five English reading time corpora: the Natural Stories [7] and Brown [8] corpora, containing self-paced reading times; and the Dundee [9], Provo [10], and GECO [11] corpora, containing first-pass (FP) and go-past (GP) durations from eye tracking. Baseline predictors in the regression 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). Per-subject random slopes were initially included for all predictors, but some were removed to ensure convergence. For each corpus and response type, the increase in log likelihood (ΔLogLik) was calculated between a regression model containing only the baseline predictors, and a regression model additionally containing an entropy predictor (either first-token entropy or MC-based word entropy). GPT2-small was the LM used to calculate entropy and surprisal predictors. MC estimates of word entropy were based on 512 next-word samples.

Results. Replacing first-token entropy with word entropy improved ΔLogLik scores in the two self-paced reading corpora, although most eye-tracking corpora showed an opposite pattern (Table 1). To evaluate the whether the observed differences were significant, a permutation test was conducted over squared errors aggregated over all corpora; this showed a significant improvement ($p < 0.01$) in ΔLogLik from word entropy compared to first-token entropy. To illustrate the difference between the two entropy predictors, Figure 1 compares the average first-token and word entropies of words in the 10 most frequent part-of-speech categories in Natural Stories. As expected, word entropy values are generally higher than first-token entropies, but the relative difference is greatest for nouns and adjectives (NN, NNS, and JJ), perhaps reflecting a wider range of multi-token words within these open-class categories.

Discussion. The results suggest that LM-based predictors operating at the word level provide a closer match to human reading times than token-level predictors, even when the former can only be approximated based on sampling. The concrete difference across the two conditions warrants caution against using first-token entropy in psycholinguistic modeling.

	Natural Stories	Brown	Dundee		Provo		GECO	
Entropy variant	SPR	SPR	FP	GP	FP	GP	FP	GP
First-token entropy	29	1.4	1.3	1.2	−0.2	1.6	3.6	−0.2
Word entropy	72	7.9	0.6	0.0	0.7	1.1	−0.4	−0.4

Table 1: Increases in log likelihood from adding the target entropy predictor to a baseline regression model for predicting self-paced reading (SPR) time, first-pass (FP) duration, or go-past (GP) duration.

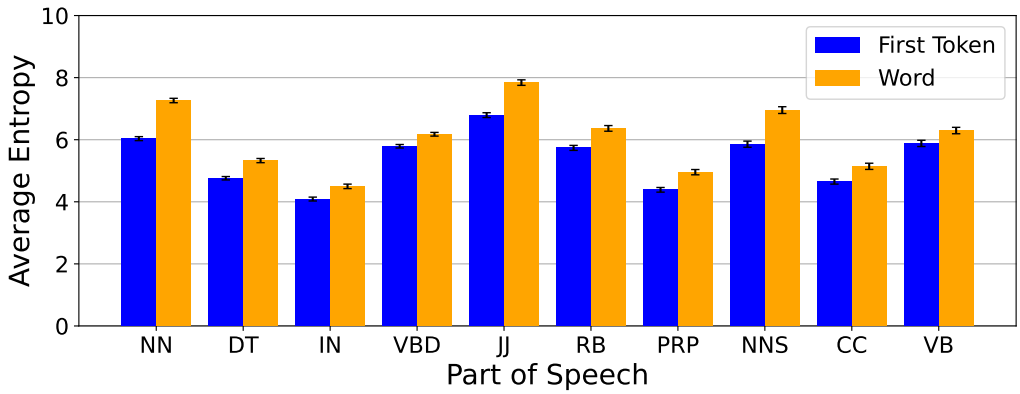

Figure 1: Average first-token and word entropy of the 10 most frequent parts of speech in the Natural Stories corpus. Part-of-speech tags are from Penn Treebank annotations [12]. Error bars represent ±1 standard error of the mean (SEM).

References

[1] Pimentel, T., et al. (2023). On the effect of anticipation on reading times. *TACL*.

[2] Wilcox, E. G., et al. (2023). Testing the predictions of surprisal theory in 11 languages. *TACL*.

[3] Giulianelli, M., et al. (2024). Generalized measures of anticipation and responsivity in online language processing. *EMNLP Findings*.

[4] Cevoli, B., et al. (2022). Prediction as a basis for skilled reading: Insights from modern language models. *Royal Society Open Science*.

[5] Radford, A., et al. (2019). Language models are unsupervised multitask learners. *ArXiv*.

[6] Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *JASA*.

[7] Futrell, R., et al. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *LREC*.

[8] Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*.

[9] Kennedy, A., et al. (2003). The Dundee corpus. *Proc. ECEM*.

[10] Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*.

[11] Cop, U., et al. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*.

[12] Marcus, M. P., et al. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*.