Unveiling the Influence of Part-of-Speech on Word Reading Time

Jeremy Xiang

Inspirit AI Research Program — 2023

Abstract

This study examines how a word's part of speech influences the time it takes to read that word. Drawing on the Provo Corpus—a collection of 55 natural paragraphs paired with eye-tracking and cloze predictability norms—we fit two ordinary least squares regression models to total reading time (TRT). Model 1 used word position, word length, log word frequency, cloze surprisal, and GPT-J surprisal as predictors. Model 2 extended this specification by adding a three-way interaction between log frequency, open-class POS membership, and GPT-J surprisal. Model 1 achieved $R^2 = 0.686$ ($F = 93.29$), and Model 2 improved to $R^2 = 0.722$ ($F = 93.73$), with the interaction term significant and large in magnitude. Word length, cloze surprisal, and log frequency were all significant predictors of reading time in the expected directions. The significance of the three-way interaction suggests that the effect of word frequency on reading time depends jointly on whether a word is open-class and on its contextual predictability as estimated by a large language model. These findings support a view of reading time as a product of interacting lexical, syntactic, and predictive factors, rather than any single feature in isolation.

1. Introduction

The time a reader spends on a single word reflects a rich set of cognitive processes: word recognition, grammatical parsing, contextual integration, and prediction of what comes next. A long-standing question in psycholinguistics is how much of this variation in reading time can be attributed to the word's part of speech (POS), independent of other features such as length, frequency, or predictability.

This paper investigates that question using a supervised regression framework. The dependent variable is total reading time (TRT) per word, measured via eye-tracking. The independent variables are a mix of numerical features (word length, log frequency, cloze probability, language-model surprisal, position in sentence) and a categorical feature (open-class vs. closed-class POS membership). The output is a quantitative estimate of how each feature, alone and in combination, predicts reading time.

Understanding the relative contribution of POS to reading time has practical implications. It informs models of reading used in education, where identifying which words are likely to slow a reader down can guide instructional pacing. It also connects to computational linguistics, where accurate reading-time predictions can serve as cognitive benchmarks for language models.

2. Background

The CMCL 2022 Shared Task on Multilingual and Crosslingual Prediction of Human Reading Behavior (Hollenstein et al., 2022) offers the most relevant precedent for this work. The task asked participants to predict five eye-tracking features—including first fixation duration and total reading time—across six languages, with a focus on features that generalize across linguistic boundaries.

Several submissions to the task used transformer-based surprisal as a feature, while others relied more heavily on classical lexical features such as word length and frequency. The best-performing models

typically combined both approaches: language model surprisal captured contextual predictability, while lexical features captured properties intrinsic to the word itself. The present work follows this combined strategy, using GPT-J surprisal alongside word length and log frequency.

The shared task's main limitation is the relatively small size of the training data (approximately 2,700 words across participants), which constrains how much model complexity is justified. This limitation also applies to the present study, which motivates the choice of ordinary least squares regression over deeper architectures: with limited data, an interpretable linear model that isolates each feature's contribution is more informative than a black-box model that is difficult to audit.

3. Dataset

The data for this study come from the Provo Corpus (Luke and Christianson, 2016, 2018), a publicly available corpus of eye-tracking data collected from 84 native English speakers reading 55 short paragraphs, each approximately 50 words long. The paragraphs were drawn from published essays, online articles, and science writing, providing a mix of registers and topics. In total, the corpus contains approximately 2,700 word tokens.

Provo was chosen over other eye-tracking corpora because it includes both eye-tracking measurements and per-word cloze predictability norms. The cloze norms were collected by asking a separate group of participants to guess the next word in each sentence, providing a human-derived measure of predictability that is rare in comparable datasets.

3.1. Preprocessing

The Provo Corpus ships with automatic POS tags in both fine-grained (Penn Treebank) and simplified (universal) form. During initial data exploration, several text-encoding issues were identified that affected tagging accuracy:

First, a small number of words were misspelled in the source text, leading the tagger to assign incorrect categories. Second, apostrophes had in some cases been replaced by question marks during file transfer, breaking contractions. Third, em-dashes were occasionally rendered as hyphens, causing adjacent words to be parsed as compounds (for example, "livres--a" and "profession--writing"). Each of these issues was corrected manually, and a cleaned version of the dataset was produced for all subsequent analyses.

3.2. Features

Each word token in the cleaned dataset was associated with the following features:

Word position (wn): the word's index within its sentence. Used to capture effects of sentence progression on reading time.

Word length: the number of characters in the word. Longer words are generally expected to require more fixation time.

Log frequency (logCount): the log of the word's frequency in the Google Web Trillion Word Corpus. Higher-frequency words are recognized more quickly.

Cloze surprisal: derived from Provo's cloze norms; higher values indicate less predictable words.

GPT-J surprisal: negative log probability assigned to the word by the GPT-J language model given the preceding context. This serves as a computational analog of human predictability.

Openness: a binary indicator of whether the word belongs to an open POS class (nouns, verbs, adjectives, adverbs) or a closed class (pronouns, prepositions, conjunctions, determiners). Open-class words carry most of the semantic content of a sentence.

The dependent variable, **TRT** (total reading time), is the summed duration of all fixations on a word across readers, provided by Provo and averaged per word.

4. Methodology

4.1. Feature Extraction

Part-of-speech tags were obtained from two sources: the Stanford CoreNLP tagger and NLTK's default tagger. Where the two taggers disagreed, the Stanford tag was used. The fine-grained Penn Treebank tags were then aggregated into higher-level categories—nouns, verbs, adjectives, adverbs, and function words—and further collapsed into a binary open-class / closed-class distinction for the regression analyses.

Word frequencies were drawn from the Google Web Trillion Word Corpus, giving a broader estimate of everyday word usage than smaller reference corpora would allow. Log frequencies were computed to normalize the heavy right-skew characteristic of word-frequency distributions.

4.2. Regression Models

Two ordinary least squares (OLS) regression models were fit using the *statsmodels* library in Python. Both models took TRT as the dependent variable.

Model 1 included five predictors: word position, word length, log frequency, cloze surprisal, and GPT-J surprisal. This model served as a baseline reflecting the standard set of features used in reading-time research.

Model 2 extended Model 1 by adding a three-way interaction term: log frequency × openness × GPT-J surprisal. This term was motivated by the hypothesis that the effect of frequency on reading time is not constant, but instead depends on whether the word carries semantic content (open-class) and on how predictable it is in context.

4.3. Evaluation

Models were compared using R^2, the F-statistic, and the Akaike Information Criterion (AIC). AIC was used as the primary model selection criterion because it penalizes model complexity in addition to rewarding fit, which is appropriate when comparing nested specifications. Individual coefficients were evaluated for statistical significance using t-tests, with 95% confidence intervals computed for visualization.

5. Results and Discussion

Both models were statistically significant as wholes. Model 1 achieved $R^2 = 0.686$ with $F = 93.29$, meaning the five baseline predictors together explain roughly 69% of the variance in TRT. Model 2 improved this to $R^2 = 0.722$ with $F = 93.73$, an increase of approximately 3.6 percentage points of explained variance. The AIC favored Model 2, indicating that the added interaction term was worth its additional complexity.

Examining the individual coefficients of Model 2 (the preferred model), nine of the ten predictors were statistically significant at the 95% confidence level. The exception was the simple interaction between open-class membership and GPT-J surprisal; this two-way term lost significance once the three-way interaction involving log frequency was introduced.

Word length had the largest effect size among the non-interaction terms, and its coefficient was positive. This matches the straightforward intuition that longer words take longer to read.

Cloze surprisal was significant and positive, but with a relatively small effect size. Less predictable words took longer to read, consistent with the long-standing finding that predictability facilitates word recognition.

Word position was significant and negative, meaning that participants read faster as they progressed through a sentence. One plausible explanation is that readers accumulate context as the sentence unfolds, which reduces the amount of integration work required for later words.

Log frequency × GPT-J surprisal was significant, large, and negative. This indicates that the slowing effect of an unexpected word (high surprisal) is attenuated when the word is itself a frequent word—a frequent but surprising word is read more quickly than a rare and surprising one.

Log frequency × openness × GPT-J surprisal was significant, large, and positive. This three-way interaction suggests that for open-class words, the interaction between frequency and surprisal is different than for closed-class words. Because open-class words carry semantic content, their frequency and predictability interact in a way that closed-class function words—which are processed more automatically—do not show.

Taken together, these results suggest that reading time cannot be explained by any single feature in isolation. The best-fitting model requires a layered account: lexical properties (length, frequency), syntactic structure (POS class), contextual predictability (cloze and GPT-J surprisal), and the interactions among them all contribute.

6. Conclusion

This study used regression analysis of the Provo Corpus to quantify how part of speech, alongside lexical and contextual features, predicts total reading time. The best-fitting model, which incorporated a three-way interaction between log frequency, open-class membership, and GPT-J surprisal, explained 72% of the variance in reading time.

Word length, cloze surprisal, and word position were all significant predictors in their expected directions. The key finding, however, was the significance of the three-way interaction: the effect of word frequency on reading time depends jointly on whether a word is open-class and on how predictable it is in context. This layered interaction is consistent with psycholinguistic theories that view reading as a process of integrating multiple information sources in parallel.

Several limitations should be noted. The Provo Corpus, while rich, is modest in size (approximately 2,700 tokens) and drawn from a limited range of written registers. Generalization to spoken language, to other genres, or to non-native readers would require additional data. The linear regression framework also assumes that effects combine additively on the scale of the dependent variable, which may understate threshold or nonlinear effects.

Future work could address these limitations by applying the same analysis to larger multilingual corpora such as the CMCL shared task data, by comparing ordinary least squares with mixed-effects models that account for between-reader variation, and by replacing GPT-J with newer language models whose surprisal estimates may correlate more closely with human reading behavior.

Acknowledgements

This paper was prepared as part of the Inspirit AI Research Program. The author thanks Professor Clayton Greenberg (University of Pennsylvania) for guidance on the research design and statistical methodology, and the Inspirit AI instructors for their ongoing mentorship throughout the project.

References

Hollenstein, N., Chersoni, E., Jacobs, C., Oseki, Y., Prévot, L., & Santus, E. (2022). CMCL 2022 Shared Task on Multilingual and Crosslingual Prediction of Human Reading Behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 121–129, Dublin, Ireland. Association for Computational Linguistics.

Luke, S. G., & Christianson, K. (2016). Limits of lexical prediction during reading. *Cognitive Psychology*, 88, 22–60.

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