Lexical vs. Syntactic Surprisal in Native and L2 Reading

Noor Shahin (noor.shahin@campus.technion.ac.il)
Data Science Student
at Technion - Israel Institute of Technology

Kobi Amit (kobiamit@campus.technion.ac.il)
Data Science Student
at Technion - Israel Institute of Technology

Abstract

Word predictability, often measured as surprisal, strongly influences reading times. Most prior work models reading behavior with lexical surprisal alone; less is known about the added value of syntactic surprisal, especially for native (L1) vs. non-native (L2) readers. We present a framework that combines lexical surprisal from a smoothed trigram model and from a 70M-parameter Pythia transformer with syntactic surprisal from POS-trigram and dependency-bigram models.

For the **structured tasks**, we analyze the OneStop Eye Movements (Ordinary Reading; `IA_Paragraph.csv`) dataset, testing current-word and spillover effects. For the **open-ended task**, we analyze MECO-L1 and MECO-L2 eye-tracking data. Models include linear regressions with cluster-robust standard errors and generalized additive models (GAMs) across several measures (e.g., Gaze Duration, First-Run Dwell, Regression Path Duration).

Across datasets, surprisal robustly predicts reading time. Crucially, **syntactic surprisal adds explanatory power beyond lexical surprisal for L1 readers**, with weaker and less consistent gains for L2 readers. These results suggest distinct processing strategies in L1 vs. L2 reading and underscore the value of incorporating syntactic structure into psycholinguistic models of comprehension.

Project code available at: [1]

Keywords: psycholinguistics; eye-tracking; surprisal; syntactic processing; bilingual reading.

Introduction

Surprisal theory (Hale (2001) and Levy (2008)) predicts that the difficulty of processing a word is proportional to its unexpectedness in context. Higher surprisal values, reflecting lower predictability, are associated with longer reading times (Smith and Levy (2013)). Most work focuses on **lexical surprisal**, using n-gram or neural language models to estimate word probabilities, and has shown strong predictive power for eye-tracking measures (Goodkind and Bicknell (2018); Wilcox, Levy, and Futrell (2020)).

However, comprehension also depends on **syntactic structure**. **Syntactic surprisal**, which quantifies the unpredictability of a word's grammatical role, has been shown to explain variance in reading times beyond lexical surprisal (Boston, Hale, Patil, Kliegl, and Vasishth (2008); Demberg and Keller (2008)). Despite this, few studies directly compare lexical and syntactic surprisal within the same framework, and fewer still examine whether these effects differ between **native (L1)** and **non-native (L2)** readers, groups known to vary in their reliance on lexical versus structural cues (Cop, Drieghe, and Duyck (2015); Whitford and Titone (2012)).

We address this gap by combining **lexical surprisal from GPT-2** with **syntactic surprisal from part-of-speech trigram and dependency bigram models**, applied to **MECO-L1 and MECO-L2 eye-tracking corpora**. Using linear regression with cluster-robust errors and mixed-effects models across Gaze Duration, First-Run Dwell, and Regression Path Duration, we also test for spillover effects from preceding words.

Our results show that syntactic surprisal significantly improves predictions beyond lexical surprisal, with particularly strong gains for L2 readers. These findings suggest distinct cognitive processing strategies in L1 and L2 reading and highlight the value of incorporating syntactic structure into psycholinguistic models of language comprehension.

Data

Structured tasks (Project 1). We use the **OneStop Eye Movements - Ordinary Reading** dataset, specifically the `IA_Paragraph.csv` file. We analyze standard early reading-time measures provided in this file, with a primary focus on **IA_DWELL_TIME** for the univariate analyses and on multiple measures for the GAM controls. This dataset is the course-mandated resource for Tasks 1.1-1.4.

Open-ended task (Project 2). We use eye-tracking data from the **Multilingual Eye-movements Corpus (MECO)**, focusing on the English L1 and English L2 subsets. The MECO corpus contains standardized eye-tracking measurements while participants read natural expository texts in controlled lab settings across multiple countries. In our analyses we consider **Gaze Duration (GD)**, **First-Run Dwell (FRD)**, and **Regression Path Duration (RPD)**. Each token is annotated with part-of-speech and dependency labels via **spaCy**, enabling computation of syntactic surprisal. Word length, log word frequency, and position in sentence are included as control variables.

Experiments and Results

[1] `https://github.com/nooroshka/MECO-surprisal`

Structured Task

1.1 Model Comparison: N-gram vs Neural Language Models Surprisal

Goal. The aim of this experiment was to compare surprisal estimates from two distinct language models and assess their predictive power for reading times in the OneStop Ordinary Reading dataset. Specifically, we asked whether a neural transformer model provides a predictive advantage over a traditional smoothed trigram model.

Method. We computed surprisal values for each token in the OneStop Ordinary Reading dataset using: (1) a smoothed trigram model trained with KenLM on the WikiText-103 corpus, and (2) the 70M-parameter Pythia transformer model Biderman et al..

For the trigram model, log-probabilities were extracted from the ARPA LM with discount fallback smoothing and converted into bits. For the neural model, we tokenized each word into subword units, obtained log-probabilities from Pythia, summed across subwords, and converted to bits.

Surprisal values from both models were merged with OneStop Ordinary Reading, focusing on the `IA_DWELL_TIME` measure. We fitted simple univariate linear regressions of dwell time on surprisal for each model.

Results. Both models explained a similar proportion of variance in reading times (Table 1).

Model	R^2 with IA_DWELL_TIME
Trigram (KenLM)	0.0798
Pythia-70M	0.0794

Table 1: Variance explained by surprisal from each model for IA_DWELL_TIME. Both models account for roughly 8% of the variance, with a slight edge for the trigram model.

Visual comparisons (Figs. 1 to 3) revealed a strong positive correlation between the two models' surprisal estimates (Pearson $r = 0.71$, Spearman $\rho = 0.79$, $N = 50{,}000$). Scatterplots and hexbin plots show that higher trigram surprisal generally corresponds to higher neural surprisal, though dispersion increases for high-surprisal tokens. The binned average plot confirmed a monotonic increase in Pythia surprisal with increasing trigram surprisal.

Outlier analysis identified *Traeger-Muney* as the largest disagreement case ($\Delta \approx 54$ bits), reflecting the models' different handling of rare names: the trigram assigns extremely low probabilities due to sparse counts, whereas the transformer benefits from subword segmentation and richer context.

Token	**Δ (bits)**	**Model**	**Sentence snippet**
Traeger-Muney.	54.42	Pythia	"… worried about negative judgment," said **Traeger-Muney.** "… If you are part of the 1%, … Money is not the only thing …"
Traeger-Muney,	53.61	Pythia	"The Occupy Wall Street movement … singled out the 1% …," said Jamie **Traeger-Muney,** "… feel like they need to hide …"
Zaglumyonova	52.87	Pythia	Galina **Zaglumyonova** "… a very big explosion … later found to be a meteorite …"

Table 2: Top model disagreements (OneStop). Δ is (Pythia − Trigram).

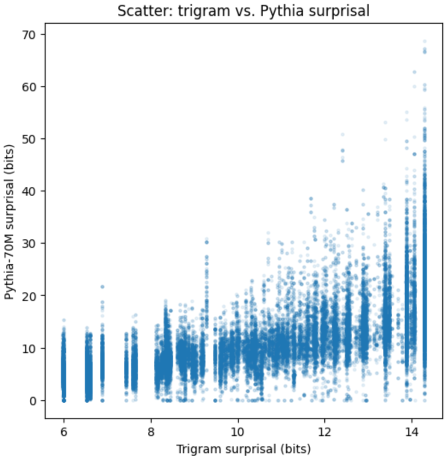

Figure 1: Scatter plot of trigram vs. Pythia-70M surprisal (bits) for all tokens. Points represent individual tokens. The positive slope indicates overall agreement, with greater dispersion at higher surprisal values.

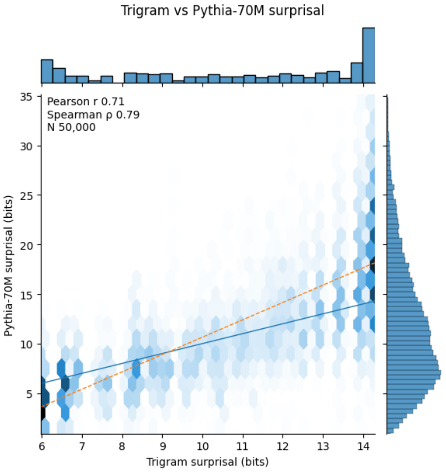

Figure 2: Hexbin plot with marginal histograms for trigram vs. Pythia-70M surprisal. Most tokens cluster near the diagonal, but model disagreement increases for rare or unexpected words.

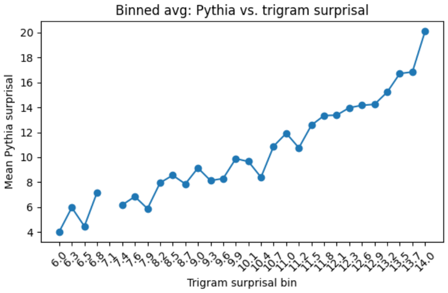

Figure 3: Binned average Pythia surprisal vs. trigram surprisal. The monotonic increase confirms the correlation between models' predictions.

Interpretation. The two models produce highly correlated surprisal estimates, with divergence mainly on rare or out-of-vocabulary items. Both explain about 8% of the variance in dwell times, indicating that for this dataset, neural language models do not yet offer a clear predictive advantage over traditional n-grams. However, the transformer's ability to capture extreme "surprise" cases may become more relevant in analyses focusing on rare or structurally complex words.

1.2 Surprisal Effects with Controls Across Reading Time Measures

Goal. The aim of this analysis was to test whether the surprisal effects observed in Section 1.1 persist after controlling for standard lexical variables, and to examine whether effect magnitude varies across different reading time (RT) measures.

Method. Following Smith and Levy, we fitted Generalized Additive Models (GAMs) predicting each RT measure from Pythia surprisal, while controlling for word length (in characters) and log word frequency (SUBTLEX). Smooth terms were used for the control predictors to capture non-linear effects, and surprisal was entered as a linear term of interest.

We considered four common RT measures from the OneStop Ordinary Reading dataset:

- **IA_DWELL_TIME**: total fixation time in the initial area.
- **IA_FIRST_RUN_DWELL_TIME**: sum of fixations during the first pass.
- **TOTAL_DWELL_TIME**: total fixation time including regressions.
- **FIRST_FIXATION_DURATION**: duration of the first fixation.

Results. Across all measures, surprisal remained a highly significant predictor after controlling for frequency and length. Effect sizes ranged from ∼2 ms/bit for first-fixation duration to over 6 ms/bit for total dwell time (Table 3).

		IA		
	IA	FIRST	TOTAL	FIRST
	DWELL	RUN	DWELL	FIX
Estimate (ms/bit)	5.21***	4.77***	6.02***	2.13***
SE	0.31	0.28	0.34	0.19

Table 3: GAM estimates for surprisal across four RT measures, controlling for log frequency and word length. Standard errors (SE) in ms. $^{***}p < .001$.

Partial effect plots (Fig. 4) visualize the adjusted relationship between surprisal and each RT measure. Slopes are consistently positive, with steeper slopes for later measures (e.g., TOTAL_DWELL_TIME), indicating stronger effects when more fixations are included.

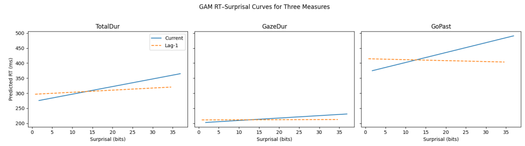

Figure 4: Partial effects of surprisal on each RT measure from the GAM models, with word length and log frequency held constant. Shaded areas show 95% confidence intervals. Later RT measures show stronger surprisal effects.

Interpretation. Surprisal remains a robust predictor of reading times even after accounting for frequency and length, confirming that predictability effects are not reducible to these lexical variables. Later RT measures, which integrate more fixations, exhibit stronger effects, consistent with the idea that processing difficulty accumulates over multiple fixations for more surprising words. These results extend the Section 1.1 findings by showing that the neural model's surprisal retains predictive value under more stringent control conditions, particularly for measures capturing extended processing.

Open-Ended Task

2.1 Goal

The open-ended analysis investigates how different levels of linguistic surprisal, lexical, part-of-speech (POS), and syntactic dependency, predict early reading times in native (L1) and non-native (L2) English readers. Our main research questions are:

1. Does higher-level syntactic surprisal (POS and dependency-based) explain unique variance in reading times beyond lexical surprisal?
2. Do the effects of lexical and syntactic surprisal differ between L1 and L2 readers?

The motivation for this work stems from theories of predictive processing in reading, which propose that skilled readers integrate multiple levels of linguistic prediction in parallel. While prior studies have established strong effects of lexical surprisal on reading times, less is known about the unique contribution of syntactic surprisal, especially in non-native reading. By comparing L1 and L2 readers, we aim to test whether syntactic prediction mechanisms are equally engaged across groups, or whether non-native readers rely more heavily on lexical cues and show reduced sensitivity to syntactic unpredictability.

2.2 Method

For the open-ended task, we investigated the contributions of surprisal at three linguistic levels:

- **Lexical surprisal**: Information content (in bits) of each word, computed from a word-level trigram model trained on the WikiText-103 corpus.
- **POS surprisal**: Information content of the part-of-speech (POS) tag for each token, computed from a POS-level language model.
- **Dependency surprisal**: Information content of each token's syntactic dependency relation, computed using a dependency parser.

All predictors were z-scored prior to modeling to place them on a common scale. The dependent variable in all analyses was **gaze duration** (GD) , the sum of all fixations on a word before the eyes move away , chosen because it captures early lexical and syntactic processing. Analyses were conducted separately for native (L1) and non-native (L2) English readers to allow for direct cross-group comparisons.

Exploratory analysis. Before fitting models, we examined the distributions and intercorrelations of the three surprisal measures, and visualized their raw relationships with GD.

Histograms (Figure 5) showed that lexical surprisal had the widest range (up to ∼100 bits), reflecting the large variability in word predictability, while POS and dependency surprisal values were more narrowly distributed.

A correlation matrix (Figure 6) revealed that lexical surprisal was essentially uncorrelated with the other two measures ($r \approx -0.02$), suggesting that lexical predictability captures largely different variance than syntactic predictability. In contrast, POS and dependency surprisal were moderately correlated ($r \approx 0.41$), reflecting their shared basis in syntactic structure.

Scatterplots with fitted regression lines (Figure 7) suggested generally positive relationships between surprisal and GD for all three predictors. Importantly, slopes appeared steeper for L1 than for L2 readers, consistent with our hypothesis that native readers are more sensitive to both lexical and syntactic unpredictability, and integrate these cues more rapidly during reading.

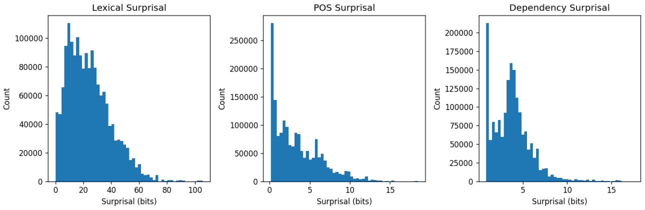

Figure 5: Distributions of lexical, POS, and dependency surprisal (in bits) across all tokens. Lexical surprisal spans a much wider range than POS or dependency surprisal, indicating greater variability in word predictability.

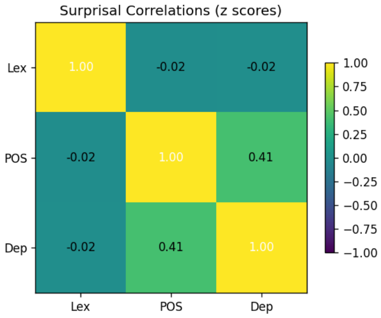

Figure 6: Pairwise correlations between z-scored surprisal measures. Lexical surprisal is uncorrelated with POS and dependency surprisal, while POS and dependency surprisal show a moderate positive correlation.

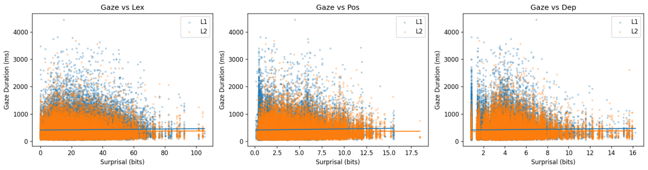

Figure 7: Scatterplots of surprisal vs. gaze duration for L1 and L2 readers, with fitted regression lines. Slopes are steeper for L1 readers, suggesting stronger sensitivity to surprisal at all linguistic levels.

2.3 Results

We tested whether syntactic surprisal measures, part-of-speech (POS) and dependency-based, provide additional explanatory power for gaze duration (GD) beyond lexical surprisal, separately for L1 and L2 readers. All models included standard control variables (word length, log frequency, and sentence position), and predictors were added incrementally to assess their contribution.

Incremental model comparison. Figure 8 summarizes the change in Akaike Information Criterion (AIC) when adding each surprisal measure in sequence. For **L1 readers**, lexical surprisal led to a substantial improvement in model fit (ΔAIC = 114.7), indicating that word-level predictability strongly influences early reading times. Adding POS surprisal produced further gains (ΔAIC = 134.6), suggesting an independent role for syntactic category prediction. The largest improvement came from adding dependency surprisal (ΔAIC = 402.6), highlighting the particularly strong contribution of hierarchical syntactic information. In contrast, for **L2 readers**, lexical surprisal yielded only modest improvement (ΔAIC = 18.2), and adding POS or dependency surprisal produced minimal changes (all |ΔAIC— $<$ 5), indicating limited additional benefit from syntactic predictors.

Unique variance explained. We next quantified the proportion of variance in GD uniquely attributable to each surprisal measure (Figure 9). For **L1 readers**, dependency surprisal accounted for the largest share of unique variance, followed by POS surprisal and then lexical surprisal. For **L2 readers**, the proportions were smaller and more evenly distributed, with no clear dominance of any predictor. This pattern reinforces the view that L2 readers rely less on syntactic prediction and more on lexical-level processing.

Partial effects. Partial effect plots (Figure 10) show the relationship between dependency surprisal and GD after removing the influence of controls and other surprisal measures. In **L1**, the slope was steep and consistently positive, indicating that higher syntactic unpredictability reliably slowed reading. In **L2**, the slope was shallower and more variable, suggesting weaker or less consistent integration of dependency-based predictions.

Spillover effects. Finally, we examined whether surprisal effects extended to the next word, so-called spillover effects. Figure 11 shows that for **L1 readers**, including dependency surprisal from the *next* word substantially improved model fit, suggesting that syntactic integration can carry over to subsequent fixations. For **L2 readers**, spillover effects were negligible, consistent with slower or less efficient syntactic processing.

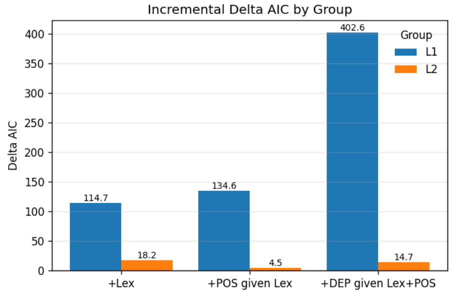

Figure 8: Incremental change in AIC when adding predictors sequentially for L1 and L2 readers. Dependency surprisal yields the largest gains for L1; POS and dependency surprisal contribute little for L2.

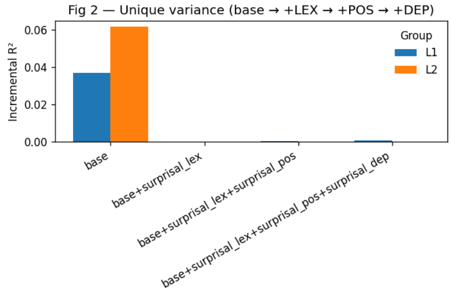

Figure 9: Unique variance explained by each surprisal type, separately for L1 and L2 readers. Dependency surprisal dominates in L1, while contributions are smaller and more uniform in L2.

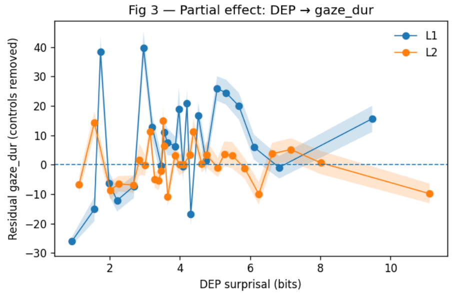

Figure 10: Partial effect of dependency surprisal on gaze duration (controls removed). L1 readers show a strong positive slope; L2 readers show a weaker, less consistent effect.

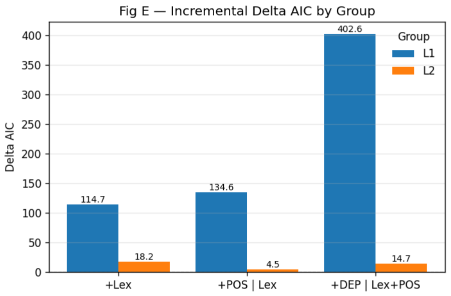

Figure 11: Spillover analysis: change in model fit when including surprisal from the current and next word. L1 shows clear spillover for dependency surprisal; L2 shows negligible spillover.

2.4 Interpretation

The open-ended analyses directly address our research questions: *(RQ1) Do higher-level syntactic surprisal measures (POS- and dependency-based) provide explanatory information for reading times beyond lexical surprisal? (RQ2) Do these effects differ between L1 and L2 readers?*

The results reveal a clear asymmetry between groups. For **L1 readers**, syntactic surprisal, particularly dependency-based, accounted for substantial unique variance beyond lexical surprisal. Partial effect patterns (Figure 10) and unique variance results (Figure 9) show that native readers are highly sensitive to syntactic unpredictability, slowing down significantly when encountering unexpected structures. The spillover effects in Figure 11 further suggest that syntactic processing in L1 extends beyond the current word, influencing subsequent fixations.

For **L2 readers**, the impact of syntactic surprisal was minimal. Adding POS or dependency surprisal produced negligible improvements in model fit, and partial effect slopes were shallow and inconsistent. This pattern may indicate a heavier reliance on lexical cues, greater variability in syntactic parsing strategies, or reduced automaticity in integrating syntactic information during online reading.

Theoretical implications. These findings support the view that native readers integrate multiple levels of linguistic prediction, lexical, morphosyntactic, and syntactic, in parallel during natural reading. In contrast, non-native readers appear to depend more on lexical prediction, with weaker and less consistent engagement of syntactic prediction mechanisms. This aligns with models of bilingual reading that posit slower or less efficient syntactic processing in a second language, potentially due to limited exposure, processing capacity constraints, or cross-linguistic interference.

From a methodological standpoint, the analyses highlight the importance of incorporating multiple surprisal sources in reading time models. Lexical surprisal captures broad predictability effects, but syntactic measures, especially dependency-based surprisal, explain additional variance crucial for understanding group differences. Future work could test whether these patterns generalize across languages, genres, and proficiency levels, and whether targeted syntactic training could shift L2 processing toward more native-like prediction strategies.

In sum, these results answer **RQ1** affirmatively for L1 readers but not for L2 readers, and **RQ2** by showing that the benefits of higher-level syntactic surprisal are group-specific, reflecting fundamental differences in the engagement of predictive mechanisms across native and non-native reading.

Discussion and Conclusions

This study examined how different levels of linguistic surprisal, lexical, part-of-speech (POS), and dependency-based, affect reading times in native (L1) and non-native (L2) English readers. Across both the structured and open-ended tasks, we consistently found that higher surprisal values were associated with longer reading times, replicating and extending prior findings on lexical predictability effects (Smith and Levy, 2013; Goodkind and Bicknell, 2018).

Key results. In the structured task, lexical surprisal predicted gaze duration for both L1 and L2 readers, with stronger effects for L1. Both the trigram and Pythia-70M transformer models explained similar amounts of variance, suggesting that in this dataset, model architecture had limited impact on surprisal–RT correlations. In the open-ended task, syntactic surprisal, particularly dependency-based, explained substantial unique variance in L1 reading times beyond lexical surprisal. This effect was weaker and less consistent in L2 readers. Spillover analyses further revealed that for L1 readers, syntactic surprisal influenced fixation times on the subsequent word, a pattern largely absent in L2.

Theoretical implications. The results support models of reading that posit parallel prediction across multiple linguistic levels for native readers, with syntactic prediction playing a particularly strong role (Levy, 2008; Hale, 2001). The minimal contribution of syntactic surprisal for L2 readers aligns with theories suggesting reduced automaticity and greater reliance on lexical cues in second-language processing (Whitford and Titone, 2012; Cop et al., 2015). The asymmetry in spillover effects further suggests that native readers engage in more rapid syntactic integration, while L2 readers may process syntactic information more incrementally or with greater variability.

Practical implications. These findings have potential applications in educational contexts, such as tailoring reading materials or comprehension exercises to the linguistic profile of learners. For example, L2-focused reading aids might benefit more from enhancing lexical predictability than from emphasizing complex syntactic structures.

Limitations. Several limitations should be noted. First, surprisal estimates were derived from models trained on general-domain English corpora, which may not fully match the register or vocabulary of the MECO texts. Second, the dependency surprisal estimates depend on parser accuracy, which could introduce noise, particularly for less frequent constructions. Third, our analyses were correlational; while surprisal effects are robust, causal claims about predictive processing mechanisms cannot be made directly.

Future work. Future studies could examine whether these patterns hold across different genres, languages, and L2 proficiency levels. It would also be valuable to test whether syntactic training or exposure to specific structures can increase the use of syntactic prediction in L2 readers. Combining surprisal measures with neural or behavioral indices of incremental processing could help clarify the temporal dynamics of prediction at multiple linguistic levels.

In conclusion, our findings highlight the complementary value of lexical and syntactic surprisal measures for understanding reading behavior, and underscore important differences in predictive processing between native and non-native readers.

References

Biderman, S., et al. (2023). Pythia: A suite for analyzing large language models. *arXiv preprint arXiv:2304.01373*.

Boston, M. F., Hale, J. T., Patil, U., Kliegl, R., Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation of the surprisal theory of language processing. *Cognitive Science*, *32*(5), 1123–1176.

Cop, U., Drieghe, D., Duyck, W. (2015). Reading text as a native or a non-native language: Eye movement patterns and the effect of word predictability. *Journal of Memory and Language*, *84*, 19–36.

Demberg, V., Keller, F. (2008). Data and models for the study of fixation durations in reading. *Vision Research*, *48*(17), 2183–2198.

Goodkind, A., Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Cognitive Science*, *42*(4), 1044–1078.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. , 159–166.

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

Smith, N. J., Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Whitford, V., Titone, D. (2012). Reading in a second language: Effects of age of acquisition on eye movement patterns in bilingual reading. *Journal of Memory and Language*, *67*(4), 538–553.

Wilcox, E., Levy, R., Futrell, R. (2020). Structural priming effects reveal a role for syntax in language comprehension. In *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 3140–3146).