Information-Theoretic Storage Cost in Sentence Comprehension

Kohei Kajikawa[1] **Shinnosuke Isono**[2] **Ethan Gotlieb Wilcox**[1]
[1]Department of Linguistics, Georgetown University, USA
[2]National Institute for Japanese Language and Linguistics, Japan
Correspondence: kk1571@georgetown.edu

Abstract

Real-time sentence comprehension imposes a significant load on working memory, as comprehenders must maintain contextual information to anticipate future input. While measures of such load have played an important role in psycholinguistic theories, they have largely been formalized using symbolic grammars, which assign discrete, uniform costs to syntactic predictions. This study proposes a measure of processing storage cost based on an information-theoretic formalization, as the amount of information previous words carry about future context, under uncertainty. Unlike previous discrete, grammar-based metrics, this measure is continuous, probabilistic, theory-neutral, and can be estimated from pre-trained neural language models. The validity of this approach is demonstrated through three analyses in English: our measure (i) recovers well-known processing asymmetries in center embeddings and relative clauses, (ii) correlates with a grammar-based storage cost in a syntactically-annotated corpus, and (iii) predicts reading-time variance in two large-scale naturalistic datasets over and above baseline models with traditional information-based predictors. Our code is available at https://github.com/kohei-kaji/info-storage.

1 Introduction

A large body of evidence shows that language comprehension is highly incremental (Marslen-Wilson, 1973; Tanenhaus et al., 1995; Kamide et al., 2003). As the input unfolds, the parser integrates incoming words into the evolving representation while generating expectations about upcoming content. However, since working memory capacity is limited, maintaining these expectations incurs a cognitive cost: as unresolved predictions accumulate, they consume memory resources and create a processing bottleneck (Just and Carpenter, 1992). It has been argued that this bottleneck gives rise to processing difficulty, for example, in center-embedding

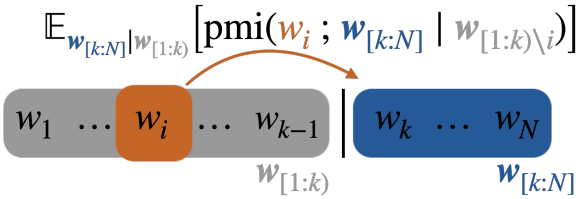


Figure 1: Illustration of the proposed storage cost measure. It quantifies the predictive potential shared between a target word w_i and the future $w_{[k:N]}$, conditioned on the remaining context. Information-theoretic storage cost at w_k is defined as the sum of the predictive potentials across all context words, representing the load of pending information.

sentences such as (1), below, where the processing of an outer clause is interrupted by an inner one, forcing the parser to hold the incomplete outer structure in working memory (Yngve, 1960; Miller and Chomsky, 1963; Gibson, 1998):

(1) The reporter [who the senator [who Mary met] attacked] ignored the president.

The memory burden of maintaining predicted syntactic elements has been discussed previously, often under the more general term of *storage cost* (Yngve, 1960; Miller and Chomsky, 1963; Abney and Johnson, 1991; Rambow and Joshi, 1994; Gibson, 1998, 2000; Kobele et al., 2013; Isono, 2024). The hypothesis that storage cost actively influences online processing has received empirical support in both controlled experiments across languages (Chen et al., 2005; Nakatani and Gibson, 2010; Stepanov and Stateva, 2015; Ristic et al., 2022) and naturalistic reading (Isono et al., 2025; Isono and Kajikawa, 2026).

Previously, most formalizations of storage costs have been derived from the incremental states of symbolic parsers (e.g., Gibson, 2000). While grammar-based storage cost successfully captures the memory burden of structural prediction, it relies on specific syntactic theories and assumes discrete,

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 15–33

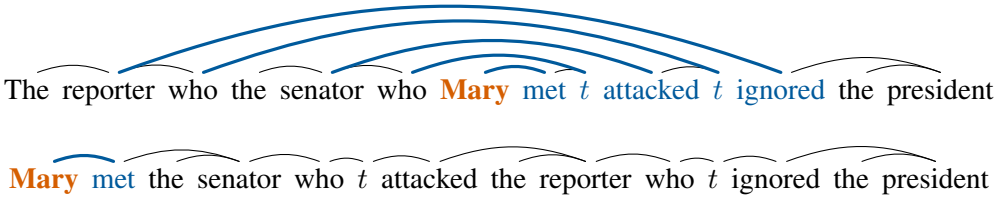


Figure 2: Illustration of DLT storage cost based on the predicted syntactic head hypothesis. In the center-embedded structure (top), the storage cost at *Mary* is **five** memory units, because five syntactic heads (the blue-colored words) are predicted to form a grammatical sentence at *Mary*. In contrast, the right-branching structure (bottom) imposes a minimal storage load. t denotes a trace of extraction. This example is adopted from Chen et al. (2005).

uniform costs for predicted elements. In this paper, we propose a grammar-independent reformulation of storage cost grounded in information theory. The key idea is that maintaining a syntactic prediction corresponds to retaining information about context words that are relevant to future input. We argue that this measure can be formalized as the extent to which the context words reduce the surprisal of future material under uncertainty. Taken in context, this is therefore the *contextualized half-pointwise mutual information* between words in context and future words. The resulting measure requires no explicit syntactic formalism, can be estimated from neural language models, and is continuous, probabilistic, and quantified in bits rather than discrete counts.

We validate the proposal through three complementary analyses. First, we examine whether the measure recovers well-known processing asymmetries in center-embedding and relative clause sentences, establishing theoretical plausibility. Second, we confirm a positive correlation with grammar-based storage cost, showing that the two measures capture overlapping structural intuitions. Third, we evaluate its predictive power on two large-scale naturalistic English reading-time datasets and compare it to that of a grammar-derived measure. We find significant improvements across multiple reading-time measures, establishing information-theoretic storage cost as a viable, theory-neutral alternative. At the same time, the variance each measure explains is largely independent, suggesting that storage cost estimated from formal grammars and neural language models accounts for partially distinct sources of reading-time variance.

2 Grammar-based Storage Cost

The idea that maintaining predicted syntactic elements consumes working memory resources has long been assumed in psycholinguistics across different grammar formalisms (Yngve, 1960; Miller and Chomsky, 1963; Abney and Johnson, 1991; Rambow and Joshi, 1994; Gibson, 1998, 2000; Kobele et al., 2013; Isono, 2024). As an example, we take Dependency Locality Theory (DLT; Gibson, 2000). Grounded in dependency grammar, DLT defines storage cost as the number of predicted syntactic heads required to complete the current input as a grammatical sentence.

Consider the following center-embedded sentence (2-a) and its right-branching variant (2-b):

(2) a. The reporter [who the senator [who **Mary** met] attacked] ignored the president.
b. Mary met the senator [who attacked the reporter [who ignored the president]].

At the point of *Mary* in (2-a), five syntactic heads are required to form a grammatical sentence: (i) three verbs to which the three subject NPs connect, and (ii) two extraction traces anticipated in object positions (see also Figure 2). Thus, DLT storage cost at *Mary* incurs a cost of five "memory units." In contrast, at each word in the right-branching variant of this sentence (2-b), at most one syntactic head is unresolved. Consequently, DLT storage cost correctly predicts the well-known difficulty of processing center-embedded structures compared to right-branching ones (Yngve, 1960; Miller and Chomsky, 1963; Gibson, 1998).

Storage cost vs. integration cost. Storage cost in DLT constitutes a complementary component to *integration cost*. While storage cost quantifies the memory burden of maintaining structural predictions before they are resolved, integration cost arises from the resource consumption required to resolve syntactic dependencies when new input is encountered. They represent different cognitive demands imposed by the same dependency structure. Integration cost has long been a focal point of

research (e.g., Gibson, 2000; Futrell et al., 2020b) because it offers a straightforward account of *locality effects*, the processing difficulty associated with longer syntactic distances (Gibson, 2000; Grodner and Gibson, 2005; Staub, 2010; Bartek et al., 2011; Roland et al., 2021). However, its empirical robustness is debated because null effects or even unexpected negative correlations (i.e., facilitation) are reported in naturalistic reading (Demberg and Keller, 2008; Shain and Schuler, 2018; Dotlačil, 2021; Isono, 2024). In this context, *lossy-context surprisal* (Futrell et al., 2020a), defined as prediction error resulting from noisy context representations, can be seen as an information-theoretic generalization of integration cost. Within this framework, locality effects emerge as a consequence of memory noise.

Linking storage cost to reading behavior. The linking hypothesis between storage cost and reading times is that storage and integration share the same pool of cognitive resources: as more resources are allocated to storage, fewer remain available for integration, resulting in slower reading times (Gibson, 1998, 2000). Controlled experiments have validated this prediction across languages, including English, Japanese, Slovenian, and Spanish (Chen et al., 2005; Nakatani and Gibson, 2010; Stepanov and Stateva, 2015; Ristic et al., 2022). Furthermore, Isono et al. (2025) demonstrate that storage costs based on both dependency grammar[1] and combinatory categorial grammar (CCG; Steedman, 2000) have larger predictive power on a held-out test set of naturalistic Japanese reading times than integration cost.

Limitations of grammar-based storage cost. Storage cost metrics based on symbolic grammars have several limitations. Theoretically, they assume that all syntactic predictions carry the same cost. Practically, it can be challenging to count the heads needed to complete a sentence. This often requires specific assumptions about grammatical completion and accurate representation of phonologically null elements (e.g., traces), which are absent in surface-based annotations like Universal Dependencies (de Marneffe et al., 2021). In addition, grammar-based metrics are often insensitive to temporal structural ambiguity, as they evaluate memory load based on a single, fully resolved parse tree. Our information-theoretic approach eliminates the need for specific syntactic theories and assumptions while capturing the probabilistic nature of sentence comprehension and replacing the discrete fixed-cost assumption with a continuous measure quantified in bits.

[1] Isono et al. (2025) adopt the Universal Dependencies framework (de Marneffe et al., 2021) and operationalize DLT storage cost as the number of unseen tokens whose co-dependents are already seen at a given word.

3 Information-Theoretic Storage Cost

This section develops an information-theoretic formalization of storage cost. We first motivate the approach and provide the formal definition, then describe how to estimate the measure using a pre-trained masked language model.

3.1 From Syntactic to Information Storage

Grammar-based storage cost counts the number of predicted syntactic elements that must be held in memory. The implicit assumption is that each prediction represents information about the future that the comprehender must maintain in working memory until it is resolved and integrated into a longer-term memory store. We propose that this can be directly quantified as the amount of information content that a word carries about future words, or the extent to which that word *reduces* the surprisal of future words. We view information-based storage as a generalization of grammar-based storage cost. Grammar-based approaches are a special case where (i) information about the future is mediated exclusively by syntactic predictions and (ii) each prediction carries a uniform cost. Our formalization relaxes both assumptions—information is measured over surface forms, and each word's contribution is a continuous quantity in bits.

3.2 Predictive Potentials as Expected PMI

We operationalize the information a word carries about the future using *contextualized pointwise mutual information* (pmi; Hoover et al., 2021). The pmi measures the association between outcomes of random variables (RVs) X and Y as the log-ratio of their joint probability to the product of their marginals:

$$\mathrm{pmi}(x; y) \coloneqq \log_2 \frac{p(X = x, Y = y)}{p(X = x)p(Y = y)} \quad (1)$$

Consider a word-valued RV W, which ranges over a vocabulary Σ, and a sentence-valued RV $\boldsymbol{W}$, which ranges over strings drawn from its Kleene closure, Σ^*. For a sentence $\boldsymbol{w}_{[1:N]} =$

$[w_1, w_2, \ldots, w_N]$, at position k, the observed context is $\boldsymbol{w}_{[1:k)} = [w_1, \ldots, w_{k-1}]$, and the future sequence is $\boldsymbol{w}_{[k:N]} = [w_k, \ldots, w_N]$. For a target word at position i (where $i < k$), we write the context excluding w_i as $\boldsymbol{w}_{[1:k)\setminus i} = [w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{k-1}]$.

The contextualized PMI between w_i and the future sequence $\boldsymbol{w}_{[k:N]}$, given the rest of the context $\boldsymbol{w}_{[1:k)\setminus i}$, is defined as:

$$\begin{aligned} &\mathrm{pmi}(w_i; \boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i}) \qquad (2)\\ &:= \log_2 \frac{p(w_i, \boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})}{p(w_i \mid \boldsymbol{w}_{[1:k)\setminus i}) p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})} \\ &= \log_2 \frac{p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)})}{p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})}. \end{aligned}$$

This quantity measures how much observing w_i changes our prediction of the future $\boldsymbol{w}_{[k:N]}$, in bits.

Since the future sequence $\boldsymbol{w}_{[k:N]}$ is unknown during left-to-right processing, we consider the expectation over possible continuations. We term this expected value *predictive potential* denoted as $\mathcal{P}_{\mathrm{pred}}$:

$$\begin{aligned} &\mathcal{P}_{\mathrm{pred}}(w_i; \boldsymbol{W}_{[k:N]}) \qquad (3)\\ &:= \mathop{\mathbb{E}}_{\boldsymbol{w}_{[k:N]} \sim p(\cdot \mid \boldsymbol{w}_{[1:k)})} \big[\mathrm{pmi}(w_i; \boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i}) \big], \end{aligned}$$

where the expectation is taken with respect to $\boldsymbol{W}_{[k:N]}$ conditioned on $\boldsymbol{w}_{[1:k)}$. This quantity is formally the *contextualized half-pointwise mutual information* and is equivalent to the Kullback-Leibler (KL) divergence between the predictive distributions with and without w_i (see Appendix A). Figure 1 provides a visual illustration of this measure.

Monotonic decay. The predictive potential of w_i regarding the future sequence is updated as processing proceeds. As the distance to the future increases, this information quantity is monotonically non-increasing in expectation (see Appendix B in detail). This property aligns with the activation decay of memory traces over time (Lewis and Vasishth, 2005; Lewis et al., 2006).

Information storage. We define the *information storage* at position k, representing the total memory load, as the sum of the predictive potentials from all preceding words:

$$\mathrm{InfoStor}_k := \sum_{i=1}^{k-1} \mathcal{P}_{\mathrm{pred}}(w_i; \boldsymbol{W}_{[k:N]}). \qquad (4)$$

It is this total information storage term that we predict is causally connected to reading difficulty.

3.3 Estimation using Masked LLMs

We estimate $\mathcal{P}_{\mathrm{pred}}$ using BERT (Devlin et al., 2019), a bidirectional masked language model. For each pair (i, k) with $i < k$, we construct inputs with and without w_i masked:

$$\begin{aligned} &\text{with } w_i: \quad \boldsymbol{w}_{[1:i)},\ w_i,\ \boldsymbol{w}_{(i:k)},\ \underbrace{\texttt{[M]}, \ldots, \texttt{[M]}}_{\boldsymbol{w}_{[k:N]}} \\ &\text{without } w_i: \quad \boldsymbol{w}_{[1:i)},\ \texttt{[M]},\ \boldsymbol{w}_{(i:k)},\ \underbrace{\texttt{[M]}, \ldots, \texttt{[M]}}_{\boldsymbol{w}_{[k:N]}} \end{aligned}$$

where `[M]` denotes a mask token and each word in $\boldsymbol{w}_{[k:N]}$ is replaced by the appropriate number of mask tokens.[2]

BERT provides token-level distributions but not joint distributions over multiple tokens. We assume conditional independence among masked positions. Let $\boldsymbol{m}$ denote the vector of masked token positions in the future region, and let $q_m^{w_i+}$ and $q_m^{w_i-}$ be BERT's predictive distributions at position m with and without w_i visible, respectively. The KL divergence then decomposes as:

$$\mathcal{P}_{\mathrm{pred}}(w_i; \boldsymbol{W}_{[k:N]}) \approx \sum_{m \in \boldsymbol{m}} D_{\mathrm{KL}}\left(q_m^{w_i+} \parallel q_m^{w_i-} \right). \qquad (5)$$

We use `bert-base-uncased` from the Transformers library (Wolf et al., 2020) in this study, in line with prior work that estimates contextualized pmi using BERT (Hoover et al., 2021; Wilcox et al., 2024).[3] We note that the assumption of conditional independence is strong (see Section 6 for future directions). We view the BERT-based methods outlined above, therefore, as a starting point which can be improved upon in future work.

4 Experiments and Results

We evaluate the information storage measure presented in Equation (4) through three analyses: (i) we visualize its distribution on specific constructions; (ii) we inspect its correlation with DLT-based storage measures; and (iii) we test its predictive power for human naturalistic reading data.

[2] We add `[CLS]` at the beginning and `[SEP]` at the end.

[3] A supplementary analysis using BERT-Large (`bert-large-uncased`) and RoBERTa (`roberta-base`; Liu et al., 2019) is provided in Appendix E. Note that the overall findings remain largely robust to the choice of model. Specifically, RoBERTa yields highly comparable results to the BERT-base model in both the correlation analysis with DLT and reading-time modeling. BERT-Large also exhibits broadly similar trends, with only minor variations.

4.1 Case Studies: Center Embeddings and Relative Clauses

To evaluate the behavior of the proposed measure, we visualize estimated storage costs for two classic syntactic asymmetries much discussed in the psycholinguistics literature. These are center-embedding vs. right-branching structures and subject vs. object relative clauses.[4] For each case, we procedurally generate sets of stimulus sentences and plot the mean information storage at each word position across these items. The full list of generated materials is provided in Appendix C.

Center embedding. Center-embedded structures (CE) incur notorious processing difficulty compared to right-branching (RB) variants (Yngve, 1960; Miller and Chomsky, 1963; Gibson, 1998). We generate 30 sentence pairs (items) using fixed templates with placeholders for animate nouns (N) and past-tense transitive verbs (V) requiring animate arguments:

CE *The* N_1 *who the* N_2 *who the* N_3 V_3 V_2 V_1 *the* N_4
RB *The* N_3 V_3 *the* N_2 *who* V_2 *the* N_1 *who* V_1 *the* N_4

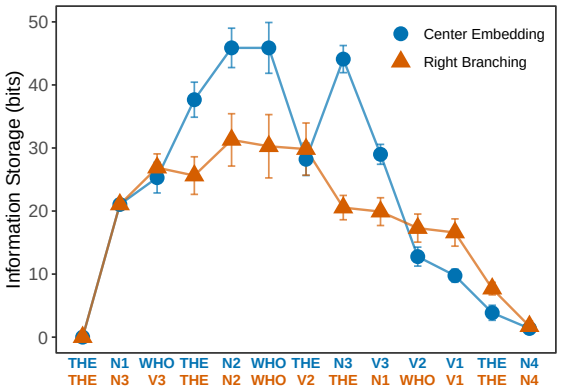
As shown in Figure 3a, information storage rises sharply in CE structures as unresolved dependencies accumulate, whereas RB structures remain lower. Summed across all words, the total information storage for CE ($\mu = 303.43$ bits, $\sigma = 44.42$) was substantially higher than for RB ($\mu = 250.54$ bits, $\sigma = 48.54$).

Subject/object relatives. Object relative clauses (ORC) are consistently more difficult to process than subject relative clauses (SRC), as evidenced by longer reading times across multiple experimental paradigms (King and Just, 1991; Grodner and Gibson, 2005; Staub, 2010; Vani et al., 2021). We generate 30 items based on the following templates:

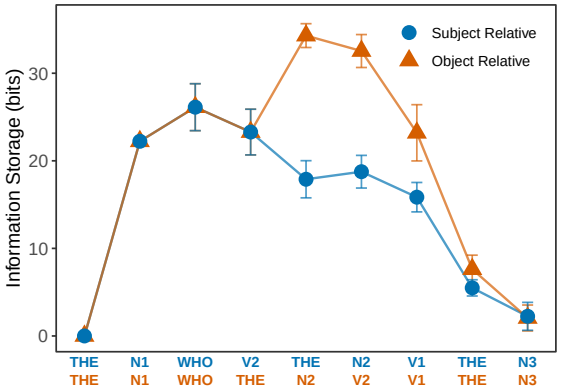
SRC *The* N_1 *who* V_2 *the* N_2 V_1 *the* N_3
ORC *The* N_1 *who the* N_2 V_2 V_1 *the* N_3

Results are visualized in Figure 3b. Our formulation predicts a critical difference between the conditions emerging at the embedded noun phrase. In ORC, the predictive potential of preceding words (e.g., *who*, *the*) accumulates, causing a peak nearly twice as high as the maximum in SRC. Consequently, ORCs yielded higher total information storage ($\mu = 171.35$ bits, $\sigma = 21.12$) than SRCs ($\mu = 131.87$ bits, $\sigma = 20.70$).

In both cases, our information-theoretic measure recovers the well-documented processing asymmetries from distributional statistics alone, without relying on syntactic annotation.



(a) Center embedding (blue) vs. right branching (orange).



(b) Subject relative (blue) vs. object relative (orange).

Figure 3: Mean information storage estimated by BERT at each word position. Error bars represent 95% confidence intervals across 30 items. In both cases, the more difficult structure (CE, ORC) exhibits higher storage cost, consistent with behavioral asymmetries.

4.2 Correlation with DLT

As a second validation of our proposal, we examine whether information storage correlates with the grammar-based DLT storage cost. We predict that correlations should be positive and moderate, as information-based storage cost contains semantic and other non-structural information not captured by purely grammar-based metrics.

[4]Note that while DLT integration cost also accounts for these asymmetries, it predicts a different locus of processing difficulty compared to storage cost. For instance, in center-embedding structures, DLT storage cost peaks at the most deeply embedded noun phrase, whereas DLT integration cost peaks at the the main verb.

We use the UD_English-GUM corpus (Zeldes, 2017), a manually annotated treebank comprising 13,263 sentences and 233,926 words. We compute both DLT storage cost and information storage for each sentence. The original definition of DLT storage cost is the number of predicted syntactic heads required to complete the current input as a grammatical sentence. Since it is challenging to apply this definition to large-scale corpora (as discussed in Section 2), we operationalize it as the count of unseen tokens whose co-dependents have been encountered. In this calculation, we exclude the following dependency relations: `punct`, `root`, `dep`, and `reparandum`. For information storage, to align BERT's subword tokenization with UD token boundaries, we define words by whitespace and sum token-level values within each word.

Figure 4 visualizes the relationship between the two measures. We observe a moderate positive correlation (Pearson's $r = 0.34$, Spearman's $\rho = 0.49$). Note, however, that the means of the DLT bins suggest that the underlying relationship may be sub-linear. This sub-linearity likely reflects a fundamental difference between the two metrics: while DLT storage increases linearly as it counts the number of unseen tokens required by the co-dependents in context based on gold parses, information storage may not do so because mutual information in natural language decays according to a power law over distance (e.g., Dębowski, 2015; Hahn et al., 2021). Regardless, the correlations suggest that our information storage captures the core intuition behind grammar-based storage cost using only distributional statistics.

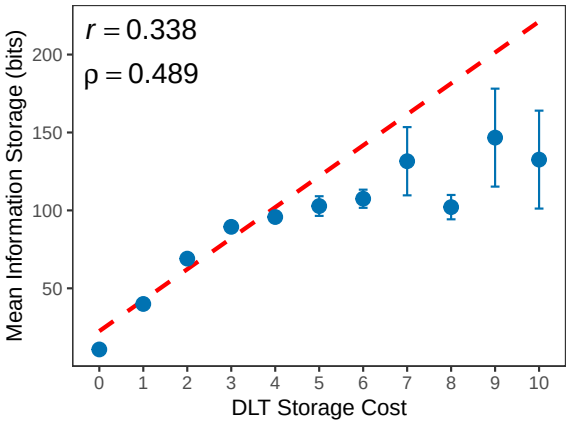


Figure 4: Mean information-theoretic storage cost as a function of DLT storage cost in the UD_English-GUM corpus (Zeldes, 2017). Points represent the mean value for each DLT bin with 95% confidence intervals. For this visualization, bins with fewer than 100 observations were excluded due to data sparsity. The red dashed line represents the linear regression fitted to the raw data corresponding to the displayed bins.

4.3 Naturalistic Reading Times

We evaluate whether information storage improves reading-time prediction in two of the largest naturalistic English reading-time datasets currently available ($N \approx 100$–200 each): Natural Stories (Futrell et al., 2021) and OneStop (Berzak et al., 2025). We also examine whether information storage explains reading-time variance above and beyond DLT storage by testing their respective contributions when the other measure is already included in the model.

4.3.1 Data

Natural Stories (Futrell et al., 2021) consists of 10 naturalistic narratives and 10,256 words. We use self-paced reading (SPR) times from 181 native English speakers (Futrell et al., 2021) and A-Maze reading times[5] from 95 native English speakers (Boyce and Levy, 2023). OneStop (Berzak et al., 2025) consists of 10 articles and 35,181 words with eye-tracking data from 180 native English speakers (the "ordinary reading" sub-portion is used). Following standard practice, we examine three eye-tracking measures: first-pass duration (FPD), the sum of all fixations from first entering a region until the first exit in either direction; go-past duration (GPD), the sum of all fixations until the first exit to the right; and total fixation duration (TFD), the sum of all fixations in a region, including re-reading.

We apply several preprocessing steps. For all datasets, we exclude the first and last words of each sentence and any words containing punctuation. For Natural Stories, we apply different criteria for each task. For self-paced reading, we follow Futrell et al. (2021) by excluding reading times shorter than 100 ms or longer than 3,000 ms, as well as participants with low comprehension accuracy (fewer than $5/6$ correct). For the A-Maze task, participants with less than 80% accuracy are excluded (Boyce and Levy, 2023).

We model the mean reading time across participants (Smith and Levy, 2013; Goodkind and

[5] A-Maze is a variant of the Maze task (Forster et al., 2009), in which, at each sentence position, participants must choose between a "true" next-word continuation and a distractor. Choice times are taken as a proxy for incremental processing times.

Bicknell, 2018; Wilcox et al., 2023a,b). Following Wilcox et al. (2023a,b), we treat skipped words in eye-tracking data as having zero ms.

4.3.2 Statistical analysis

Following standard statistical model comparison and prior work on naturalistic reading-time analysis (e.g., Frank and Bod, 2011; Goodkind and Bicknell, 2018; Wilcox et al., 2023b), we test whether adding storage cost measures to a baseline linear regression model improves reading-time prediction. The baseline model includes word positions (in the sentence and document), word length (the number of characters), unigram surprisal, and GPT-2 surprisal. DLT storage is calculated as the number of unseen tokens whose co-dependents are already seen at a given word. To obtain these dependencies, we use UD parses in Natural Stories (Futrell et al., 2021), and parses generated by `Stanza` (Qi et al., 2020) for OneStop. Unigram surprisal is estimated using the implementation by Oh et al. (2024) on approximately 33 billion pre-tokenized tokens from the Pile dataset (Gao et al., 2020). For GPT-2 surprisal, we use the 124M-parameter GPT-2 (Radford et al., 2019), as it is a better predictor of reading times than larger models (Oh and Schuler, 2023; Shain et al., 2024). To estimate GPT-2 surprisal, we use a context of up to 1,024 preceding tokens and adopt whitespace-trailing decoding (Oh and Schuler, 2024), which reassigns the probability of a leading whitespace to the preceding word. We include spillover terms for word length, unigram surprisal, GPT-2 surprisal, and both storage measures. All predictors are z-scored. Note that the storage measures are not highly correlated with other predictors. The full correlation matrix is provided in Figure 7 of Appendix D.[6]

To assess the predictive power of the storage measures for reading-time variance, we evaluate the change in predictive power under four conditions: (i) adding information storage to the baseline (INFO), (ii) adding DLT storage to the baseline (DLT), (iii) adding information storage to the baseline that already includes DLT storage (INFO-ON-DLT), and (iv) adding DLT storage to the baseline that already includes information storage (DLT-ON-INFO). Conditions INFO and DLT test the independent contribution of each measure, while INFO-ON-DLT and DLT-ON-INFO test whether each measure provides additional predictive power when the other is controlled for.

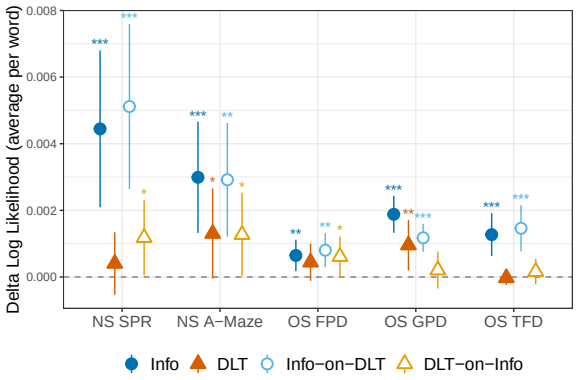
We evaluate predictive power using the per-word change in log-likelihood (Δ_{LL}) between the target and baseline models. 10-fold cross-validation (CV) is employed to estimate Δ_{LL} on held-out test data. Significance is assessed via a one-sided permutation test (20,000 iterations) using the mean Δ_{LL} as the test statistic. To account for multiple comparisons across all 20 tests (5 datasets $\times$ 4 conditions), the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) is applied to control the false discovery rate (FDR) at $\alpha = 0.05$. We also examine the mean coefficient for information storage and DLT storage (averaged across the 10 CV folds) in the INFO and DLT conditions to verify whether they are positive, as hypothesized.

4.3.3 Results

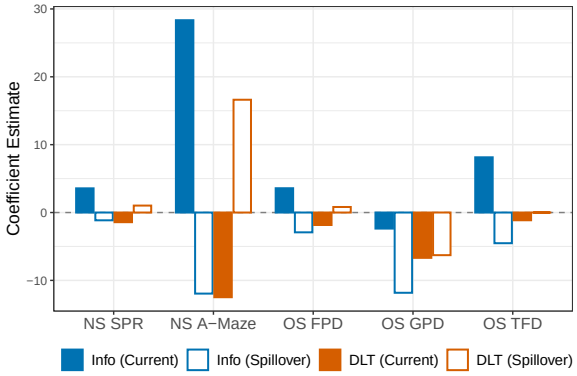
Figure 5a illustrates the predictive contributions of each storage measure. Individually, information storage (INFO) significantly improved model fit in all five datasets, while DLT storage (DLT) is significant in Natural Stories A-Maze and OneStop GPD. The above-and-beyond analysis (INFO-ON-DLT and DLT-ON-INFO conditions) further reveals that the two measures capture largely independent aspects of reading-time variance. Adding information storage to a model that already contains DLT storage always yields significant improvements, whereas adding DLT storage to a model that includes information storage results in significant improvements in three out of five tests. Crucially, the magnitude of improvement for one measure remains largely stable regardless of whether the other is included in the model (compare INFO vs. INFO-ON-DLT, and DLT vs. DLT-ON-INFO). This consistent pattern suggests that while our information-theoretic measure significantly predicts reading times, it and the grammar-based DLT are complementary, each accounting for distinct sources of reading-time variance. This is not necessarily surprising, given that LLM predictions are driven by far more than just structural information (Hu et al., 2026; McGee et al., 2026), and conversely, humans may rely more on structured information than LLMs (Kajikawa and Isono, 2026).

Figure 5b shows the mean regression coefficients

[6] As information storage exhibits high autocorrelation ($r = 0.85$ in both datasets), we compute the Variance Inflation Factor (VIF) for all predictors to ensure that multicollinearity does not destabilize the regression estimates. The VIF values peak at approximately 3.8 for information storage and its spillover term. This remains well below the conservative threshold of 5, which indicates that essentially no additional confounding collinearity exists.



(a) Predictive power of storage measures quantified by per-word Δ_{LL}. Results are shown for four model comparisons. Points and error bars represent means and 95% confidence intervals, respectively. Stars indicate FDR-adjusted significance levels (*** $q < .001$, ** $q < .01$, * $q < .05$, n.s. $q \geq 0.05$).



(b) Mean regression coefficients for information storage and DLT storage at the current and spillover regions, averaged across 10 CV folds in the INFO and DLT conditions, respectively. As the predictors were z-scored, coefficients represent relative effect sizes.

Figure 5: Results of the naturalistic reading-time analysis on Natural Stories (NS) and OneStop (OS). Abbreviations: SPR = self-paced reading; FPD = first-pass duration; GPD = go-past duration; TFD = total fixation duration.

for information and DLT storage across folds of data for our linear regression models in the INFO and DLT conditions, respectively. For information storage, coefficients in the current region are positive for all datasets except for OneStop GPD, supporting the hypothesis that higher information storage increases processing difficulty. For OneStop GPD, the coefficient is negative. This divergence between GPD and TFD suggests a specific reading strategy: high storage cost may prompt readers to move their gaze rightward more quickly to resolve uncertainty, subsequently leading to the regressions reflected in the increased total fixation duration. As for DLT storage, the coefficients in the current region are consistently negative. Recent studies on Japanese, an SOV language, suggest that readers exhibit differing processing strategies in response to high DLT storage cost, resulting in variable directions of the effect (Isono et al., 2025; Isono and Kajikawa, 2026). Against this backdrop, it remains an important question for future research to explore whether our negative effect reflects a uniform speed-up strategy among readers on average or stems from language-specific structural factors.

5 Discussion

In this study, we introduced *information storage cost*, an information-theoretic measure of working memory storage quantifying expectations about future input. Unlike traditional storage cost metrics that rely on specific syntactic theories and discrete counting, our measure is continuous, grammar-independent, and estimable from neural language models. Our analyses demonstrated the validity of this approach: information storage successfully recovers the processing difficulty of center embeddings and object relative clauses solely from distributional statistics, correlates with grammar-based DLT storage, and predicts reading times in naturalistic text over and above baseline models. These results suggest that the cognitive bottleneck of storage can be effectively operationalized as the maintenance of (half-pointwise) mutual information between the context and the future.

The current theory is noteworthy in the context of *lossy-context surprisal* (Futrell et al., 2020a). Our measure operationalizes storage cost as the sum of predictive potentials under the assumption of perfect context maintenance. It quantifies the informational load required to sustain the high-fidelity representations that underlie prediction errors in traditional surprisal theory (Hale, 2001; Levy, 2008). However, human sentence processing is fundamentally constrained by memory loss and noise (Lewis and Vasishth, 2005; Levy et al., 2009). Just as lossy-context surprisal (Futrell et al., 2020a) generalized traditional surprisal by integrating memory constraints through noisy context representations, our framework provides a foundation for extending storage cost to account for such imperfect maintenance. Thus, investigating *lossy-context* storage remains a crucial next step to more accurately capture how the cognitive system manages information under finite resources. Ultimately, while prior research on the *memory-surprisal tradeoff* (Hahn et al., 2021, 2022b) has

primarily focused on corpus-level averages, our proposed metric provides the formal machinery to investigate these information-theoretic dynamics at the level of incremental, word-by-word sentence comprehension.

The observed alignment between our information-theoretic measure and grammar-based DLT storage is theoretically linked to the Head-Dependent Mutual Information (HDMI) hypothesis proposed by Futrell et al. (2019). The HDMI hypothesis posits that syntactic dependencies exist primarily between word pairs exhibiting high PMI. Indeed, Futrell et al. (2020a) provided empirical evidence across 54 languages showing that words in a syntactic dependency share higher PMI on average than those without such a relation. This theoretical framework explains why our information-theoretic metric, despite being derived purely from distributional statistics, correlates with storage costs grounded in dependency grammar.

However, our analysis of naturalistic reading times revealed that this alignment is somewhat orthogonal to their predictive power: information storage and DLT storage capture partially distinct sources of reading-time variance. This finding implies that the information-theoretic approach is not merely a continuous generalization that subsumes discrete grammar-based metrics. Crucially, the fact that DLT storage remains a robust predictor even after controlling for information storage highlights the importance of structure-based metrics. Given the fundamental bottleneck of working memory (Christiansen and Chater, 2016), it is not surprising that sentence comprehension involves abstract structural knowledge alongside statistical patterns. To cope with the rapid loss of linguistic input, the cognitive system likely relies on symbolic structural representations to chunk or compress information efficiently—a function that probabilistic prediction based on statistical patterns alone may not fully fulfill. Thus, statistical prediction and structural processing appear to operate as distinct, complementary mechanisms in overcoming the cognitive constraints of language comprehension.

6 Limitations

This study has several limitations, which are important to acknowledge. We defined information storage at the word level, assuming that comprehenders maintain specific lexical predictions. However, cognitive resource-rationality in sentence processing suggests that memory representations are likely compressed to optimize the tradeoff between precision and capacity (Hahn et al., 2022a; Xu and Futrell, 2026). Future work should explore the optimal granularity of compressed representations for modeling storage cost, rather than relying on raw tokens.

When estimating values of information storage using BERT, we assumed conditional independence between these tokens. While this is a strong and unrealistic assumption for natural language, overcoming it within our current formulation presents a severe computational hurdle. Since our definition of predictive potential requires estimating the joint probability of the future sequence conditioned on a gapped context (see Equation (3)), evaluating this quantity with autoregressive language models is theoretically possible but computationally prohibitive, as it necessitates marginalizing over the entire vocabulary at the omitted position i at every incremental step. As a radically alternative direction, rather than engineering a workaround for this computational dilemma, one could reconsider the underlying theoretical formulation itself. Instead of aggregating the word-by-word predictive potentials, storage cost at k could be reformulated at a holistic level as the *half-pointwise predictive information* (mutual information) between the observed context and the future sequence:

$$\begin{aligned}
&\mathrm{I}(\boldsymbol{W}_{[1:k)} = \boldsymbol{w}_{[1:k)}; \boldsymbol{W}_{[k:N]}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{w}_{[k:N]} \sim p(\cdot \mid \boldsymbol{w}_{[1:k)})} \big[\operatorname{pmi}(\boldsymbol{w}_{[1:k)}; \boldsymbol{w}_{[k:N]}) \big] \\
&= D_{\mathrm{KL}} \Big(p(\boldsymbol{W}_{[k:N]} \mid \boldsymbol{w}_{[1:k)}) \parallel p(\boldsymbol{W}_{[k:N]}) \Big). \quad (6)
\end{aligned}$$

Crucially, this shift would eliminate the conditional independence assumption entirely, albeit at the expense of a distinct computational bottleneck: the need to approximate the unconditional marginal distribution of the future sequence, $p(\boldsymbol{W}_{[k:N]})$.

Finally, the interplay between storage and prediction likely varies across languages, particularly in head-final (SOV) languages like Japanese, where pre-verbal memory demands are structurally higher (Nakatani and Gibson, 2010; Isono et al., 2025; Isono and Kajikawa, 2026). Investigating these compression strategies and cross-linguistic variations offers a promising avenue for refining information-theoretic models of working memory.

Acknowledgments

We are grateful to Lin Ai and Taiga Someya for their valuable comments on earlier versions of this work. This work is supported by JSPS KAKENHI Grant Number JP25K22996.

References

Steven P. Abney and Mark Johnson. 1991. Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, 20(3):233–250.

Brian Bartek, Richard L. Lewis, Shravan Vasishth, and Mason R. Smith. 2011. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178–1198.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. 2025. OneStop: A 360-participant English eye tracking dataset with different reading regimes. *Scientific Data*.

Veronica Boyce and Roger Levy. 2023. A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics*, 2(1).

Evan Chen, Edward Gibson, and Florian Wolf. 2005. Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, 52(1):144–169.

Morten H. Christiansen and Nick Chater. 2016. The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39:e62.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Łukasz Dębowski. 2015. The relaxed Hilberg conjecture: A review and new experimental support. *Journal of Quantitative Linguistics*, 22(4):311–337.

Jakub Dotlačil. 2021. Parsing as a cue-based retrieval model. *Cognitive Science*, 45(8):e13020.

Kenneth I. Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, 41(1):163–171.

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020a. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. 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.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020b. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 3–13, Paris, France. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. The MIT Press.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, 29(2):261–290.

Michael Hahn, Judith Degen, and Richard Futrell. 2021. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*, 128:726–756.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022a. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

Michael Hahn, Rebecca Mathew, and Judith Degen. 2022b. Morpheme ordering across languages reflects optimization for processing efficiency. *Open Mind: Discoveries in Cognitive Science*, 5:208–232.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Jacob Louis Hoover, Wenyu Du, Alessandro Sordoni, and Timothy J. O'Donnell. 2021. Linguistic dependencies and statistical dependence. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2963, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jennifer Hu, Ethan Gotlieb Wilcox, Siyuan Song, Kyle Mahowald, and Roger P. Levy. 2026. What can string probability tell us about grammaticality? *Transactions of the Association for Computational Linguistics*, 14:124–146.

Shinnosuke Isono. 2024. Category locality theory: A unified account of locality effects in sentence comprehension. *Cognition*, 247:105766.

Shinnosuke Isono and Kohei Kajikawa. 2026. Syntactically-guided information maintenance in sentence comprehension. *arXiv preprint arXiv:2604.27468*.

Shinnosuke Isono, Kohei Kajikawa, Yohei Oseki, and Masayuki Asahara. 2025. Modeling memory effects in a head-final language with category locality. *PsyArXiv preprint*.

Marcel A. Just and Patricia A. Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1):122–149.

Kohei Kajikawa and Shinnosuke Isono. 2026. The dual nature of syntactic Node Count: Facilitating and inhibiting sentence comprehension. *PsyArXiv preprint*.

Yuki Kamide, Christoph Scheepers, and Gerry T. M. Altmann. 2003. Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, 32(1):37–55.

Jonathan King and Marcel Adam Just. 1991. Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5):580–602.

Gregory M. Kobele, Sabrina Gerth, and John Hale. 2013. Memory resource allocation in top-down minimalist parsing. In *Formal Grammar*, pages 32–51, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. 2009. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10:447–454.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

William Marslen-Wilson. 1973. Linguistic structure and speech shadowing at very short latencies. *Nature*, 244:522–523.

Thomas A. McGee, Yiyang Zhang, and Idan A. Blank. 2026. Evidence against syntactic encapsulation in large language models. *Cognitive Science*, 50(3):e70187.

George A. Miller and Noam Chomsky. 1963. Finitary models of language users. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, editors, *Handbook of mathematical psychology*, volume 2, pages 419–491. Wiley.

Kentaro Nakatani and Edward Gibson. 2010. An online study of Japanese nesting complexity. *Cognitive Science*, 34(1):94–112.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

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.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

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

Owen Rambow and Aravind K. Joshi. 1994. A processing model for free word-order languages. In Charles Clifton, Lyn Frazier, and Keith Rayner, editors, *Perspectives on Sentence Processing*, pages 267–301. Lawrence Erlbaum Associates, Inc.

Bojana Ristic, Simona Mancini, Nicola Molinaro, and Adrian Staub. 2022. Maintenance cost in the processing of subject–verb dependencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(6):829–838.

Douglas Roland, Gail Mauner, and Yuki Hirose. 2021. The processing of pronominal relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 119:104244.

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.

Cory Shain and William Schuler. 2018. Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2679–2689, Brussels, Belgium. Association for Computational Linguistics.

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

Adrian Staub. 2010. Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1):71–86.

Mark Steedman. 2000. *The syntactic process*. MIT Press.

Artur Stepanov and Penka Stateva. 2015. Cross-linguistic evidence for memory storage costs in filler-gap dependencies with wh-adjuncts. *Frontiers in Psychology*, 6.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Pranali Vani, Ethan Gotlieb Wilcox, and Roger Levy. 2021. Using the interpolated maze task to assess incremental processing in English relative clauses. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023a. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

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

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2024. An information-theoretic analysis of targeted regressions during reading. *Cognition*, 249:105765.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weijie Xu and Richard Futrell. 2026. Strategic resource allocation in memory encoding: An efficiency principle shaping language processing. *Journal of Memory and Language*, 146:104706.

Victor H Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A Predictive Potential as KL Divergence

Derivations establishing the equivalence of predictive potential to a KL divergence are provided.

Starting from the definition of contextualized PMI:

$$\begin{aligned}\operatorname{pmi}(w_i; \boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i}) &\coloneqq \log \frac{p(w_i, \boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})}{p(w_i \mid \boldsymbol{w}_{[1:k)\setminus i}) \cdot p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})} \\ &= \log \frac{p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i}, w_i) \cdot \cancel{p(w_i \mid \boldsymbol{w}_{[1:k)\setminus i})}}{\cancel{p(w_i \mid \boldsymbol{w}_{[1:k)\setminus i})} \cdot p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})} \\ &= \log \frac{p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)})}{p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})}. \end{aligned} \tag{7}$$

Taking the expectation over $\boldsymbol{w}_{[k:N]}$ conditioned on $\boldsymbol{w}_{[1:k)}$:

$$\begin{aligned}\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]}) &\coloneqq \mathop{\mathbb{E}}_{\boldsymbol{w}_{[k:N]} \sim p(\cdot \mid \boldsymbol{w}_{[1:k)})} \big[\operatorname{pmi}(w_i; \boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i}) \big] \\ &= \sum_{\boldsymbol{w}_{[k:N]} \in \Sigma^*} p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)}) \log \frac{p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)})}{p(\boldsymbol{w}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})} \\ &= D_{\text{KL}}\Big(p(\boldsymbol{W}_{[k:N]} \mid \boldsymbol{w}_{[1:k)}) \parallel p(\boldsymbol{W}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i}) \Big) \geq 0. \end{aligned} \tag{8}$$

Non-negativity follows from the non-negativity of KL divergence.

B Monotonic Non-Increase in Expectation

We show that the predictive potential, the contextualized half-pointwise mutual information, is monotonically non-increasing in expectation over the next word. In other words, the following relation holds:

$$\mathop{\mathbb{E}}_{w_k \sim p(\cdot \mid \boldsymbol{w}_{[1:k)})} \big[\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k+1:N]}) \big] \leq \mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]}). \tag{9}$$

Chain rule for KL divergence. We first introduce the chain rule for KL divergence. For any joint distributions $P(X, Y)$ and $Q(X, Y)$, applying the definition of KL divergence and the factorization $P(X, Y) = P(X)P(Y \mid X)$, we have:

$$\begin{aligned} &D_{\text{KL}}\Big(P(X, Y) \parallel Q(X, Y) \Big) \\ &= \sum_{x,y} P(x, y) \log \frac{P(x, y)}{Q(x, y)} \\ &= \sum_{x,y} P(x, y) \log \frac{P(x)P(y \mid x)}{Q(x)Q(y \mid x)} \\ &= \sum_{x,y} P(x, y) \log \frac{P(x)}{Q(x)} + \sum_{x,y} P(x, y) \log \frac{P(y \mid x)}{Q(y \mid x)} \\ &= D_{\text{KL}}\Big(P(X) \parallel Q(X) \Big) + \sum_{x,y} P(x)P(y \mid x) \log \frac{P(y \mid x)}{Q(y \mid x)} \\ &= D_{\text{KL}}\Big(P(X) \parallel Q(X) \Big) + \sum_{x} P(x) \sum_{y} P(y \mid x) \log \frac{P(y \mid x)}{Q(y \mid x)} \\ &= D_{\text{KL}}\Big(P(X) \parallel Q(X) \Big) + \mathbb{E}_X \Big[D_{\text{KL}}\Big(P(Y \mid X) \parallel Q(Y \mid X) \Big) \Big]. \end{aligned} \tag{10}$$

This states that the divergence between joint distributions equals the divergence between marginals plus the expected divergence between conditionals.

The predictive potential of w_i at position k is:

$$\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]}) = D_{\text{KL}}\Big(p(\boldsymbol{W}_{[k:N]} \mid \boldsymbol{w}_{[1:k)}) \parallel p(\boldsymbol{W}_{[k:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})\Big). \tag{11}$$

Similarly, at position $k+1$:

$$\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k+1:N]}) = D_{\text{KL}}\Big(p(\boldsymbol{W}_{[k+1:N]} \mid \boldsymbol{w}_{[1:k+1)}) \parallel p(\boldsymbol{W}_{[k+1:N]} \mid \boldsymbol{w}_{[1:k+1)\setminus i})\Big). \tag{12}$$

We decompose $\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]})$ using the chain rule for KL divergence (10). Note that $\boldsymbol{W}_{[k:N]} = (W_k, \boldsymbol{W}_{[k+1:N]})$. Applying the chain rule:

$$\begin{aligned}
\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]}) &= D_{\text{KL}}\Big(p(W_k, \boldsymbol{W}_{[k+1:N]} \mid \boldsymbol{w}_{[1:k)}) \parallel p(W_k, \boldsymbol{W}_{[k+1:N]} \mid \boldsymbol{w}_{[1:k)\setminus i})\Big) \\
&= D_{\text{KL}}\Big(p(W_k \mid \boldsymbol{w}_{[1:k)}) \parallel p(W_k \mid \boldsymbol{w}_{[1:k)\setminus i})\Big) \\
&+ \mathop{\mathbb{E}}_{w_k \sim p(\cdot \mid \boldsymbol{w}_{[1:k)})}\Big[D_{\text{KL}}\Big(p(\boldsymbol{W}_{[k+1:N]} \mid \boldsymbol{w}_{[1:k)}, w_k) \parallel p(\boldsymbol{W}_{[k+1:N]} \mid \boldsymbol{w}_{[1:k)\setminus i}, w_k)\Big)\Big].
\end{aligned} \tag{13}$$

Conditioning on $W_k = w_k$, we have $\boldsymbol{w}_{[1:k+1)} = (\boldsymbol{w}_{[1:k)}, w_k)$ and $\boldsymbol{w}_{[1:k+1)\setminus i} = (\boldsymbol{w}_{[1:k)\setminus i}, w_k)$. Thus, the KL divergence inside the expectation equals $\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k+1:N]})$ evaluated at context $\boldsymbol{w}_{[1:k+1)}$.

Since KL divergence is always non-negative, we obtain:

$$\mathop{\mathbb{E}}_{w_k \sim p(\cdot \mid \boldsymbol{w}_{[1:k)})}\big[\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k+1:N]})\big] \le \mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]}). \tag{14}$$

This shows that, in expectation over the word $W_k = w_k$, the predictive potential of w_i decreases (or stays the same) as the sentence proceeds. Of course, for a specific observed word w_k, the value $\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k+1:N]})$ may exceed $\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]})$, but $\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k+1:N]})$ does not exceed $\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]})$ on average.

To verify this theoretical result empirically, we approximate the expectation via a Monte Carlo estimate using the UD_English-GUM. Specifically, we compute the sample mean of predictive potentials as a function of distance $d = k - i$. For each distance d up to 30, we collect all $\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]})$.

Figure 6 shows that the mean predictive potential decreases monotonically with distance, consistent with the theoretical prediction. This decay pattern reflects the general property that words become less predictive of future materials as the distance gets longer.

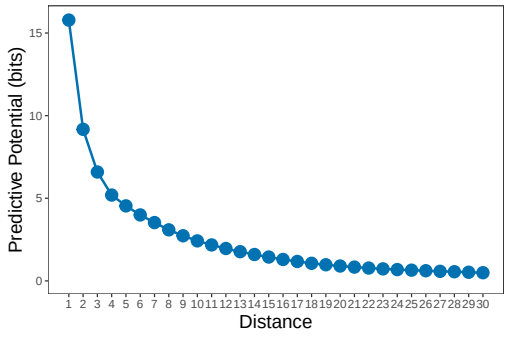


Figure 6: Mean predictive potential $\mathcal{P}_{\text{pred}}(w_i; \boldsymbol{W}_{[k:N]})$ as a function of distance $d = k - i$ (up to a maximum of 30), computed over the UD_English-GUM. The monotonic decrease is consistent with the theoretical result that predictive potential is non-increasing on average.

C Generated Stimulus Materials

The complete set of 30 items used for the illustrative analyses in Section 4.1 is listed. Each item consists of two syntactically distinct but lexically identical sentences.

Center-embedding vs. right-branching structures The following items contrast double center-embedded structures (a) with their corresponding right-branching variants (b).

(1) a. The actor who the doctor who the reporter watched visited called the dancer
b. The reporter watched the doctor who visited the actor who called the dancer
(2) a. The lawyer who the dancer who the doctor attacked stopped praised the writer
b. The doctor attacked the dancer who stopped the lawyer who praised the writer
(3) a. The judge who the pilot who the senator watched visited attacked the reporter
b. The senator watched the pilot who visited the judge who attacked the reporter
(4) a. The nurse who the thief who the reporter avoided praised visited the guard
b. The reporter avoided the thief who praised the nurse who visited the guard
(5) a. The singer who the guard who the pilot called questioned admired the student
b. The pilot called the guard who questioned the singer who admired the student
(6) a. The reporter who the artist who the singer ignored called helped the pilot
b. The singer ignored the artist who called the reporter who helped the pilot
(7) a. The banker who the detective who the doctor helped attacked warned the president
b. The doctor helped the detective who attacked the banker who warned the president
(8) a. The neighbor who the thief who the teacher stopped admired praised the senator
b. The teacher stopped the thief who admired the neighbor who praised the senator
(9) a. The doctor who the chef who the president avoided watched called the guard
b. The president avoided the chef who watched the doctor who called the guard
(10) a. The thief who the neighbor who the judge met attacked praised the banker
b. The judge met the neighbor who attacked the thief who praised the banker
(11) a. The writer who the student who the soldier attacked visited watched the president
b. The soldier attacked the student who visited the writer who watched the president
(12) a. The chef who the teacher who the baker ignored helped watched the actor
b. The baker ignored the teacher who helped the chef who watched the actor
(13) a. The neighbor who the banker who the writer watched avoided praised the teacher
b. The writer watched the banker who avoided the neighbor who praised the teacher
(14) a. The president who the thief who the actor visited praised stopped the artist
b. The actor visited the thief who praised the president who stopped the artist
(15) a. The artist who the baker who the chef praised avoided trusted the teacher
b. The chef praised the baker who avoided the artist who trusted the teacher
(16) a. The guard who the student who the writer trusted noticed watched the detective
b. The writer trusted the student who noticed the guard who watched the detective
(17) a. The senator who the student who the detective visited attacked called the chef
b. The detective visited the student who attacked the senator who called the chef
(18) a. The judge who the singer who the detective warned admired avoided the banker
b. The detective warned the singer who admired the judge who avoided the banker
(19) a. The actor who the baker who the lawyer praised visited ignored the teacher
b. The lawyer praised the baker who visited the actor who ignored the teacher
(20) a. The guard who the teacher who the dancer trusted questioned warned the judge
b. The dancer trusted the teacher who questioned the guard who warned the judge
(21) a. The neighbor who the student who the lawyer noticed attacked admired the nurse
b. The lawyer noticed the student who attacked the neighbor who admired the nurse
(22) a. The senator who the doctor who the lawyer avoided noticed ignored the actor
b. The lawyer avoided the doctor who noticed the senator who ignored the actor
(23) a. The pilot who the thief who the president admired questioned warned the chef
b. The president admired the thief who questioned the pilot who warned the chef
(24) a. The nurse who the teacher who the guard attacked praised avoided the reporter
b. The guard attacked the teacher who praised the nurse who avoided the reporter
(25) a. The writer who the guard who the teacher called attacked helped the actor
b. The teacher called the guard who attacked the writer who helped the actor
(26) a. The pilot who the artist who the baker called trusted praised the reporter
b. The baker called the artist who trusted the pilot who praised the reporter
(27) a. The nurse who the artist who the doctor avoided watched called the actor
b. The doctor avoided the artist who watched the nurse who called the actor
(28) a. The nurse who the thief who the banker ignored noticed helped the lawyer
b. The banker ignored the thief who noticed the nurse who helped the lawyer
(29) a. The guard who the nurse who the reporter met admired helped the thief
b. The reporter met the nurse who admired the guard who helped the thief
(30) a. The doctor who the neighbor who the soldier questioned visited met the student
b. The soldier questioned the neighbor who visited the doctor who met the student

Subject vs. object relatives The following items contrast subject relative clauses (a) with object relative clauses (b).

(1) a. The actor who called the doctor praised the reporter
b. The actor who the doctor called praised the reporter
(2) a. The student who avoided the lawyer attacked the dancer
b. The student who the lawyer avoided attacked the dancer

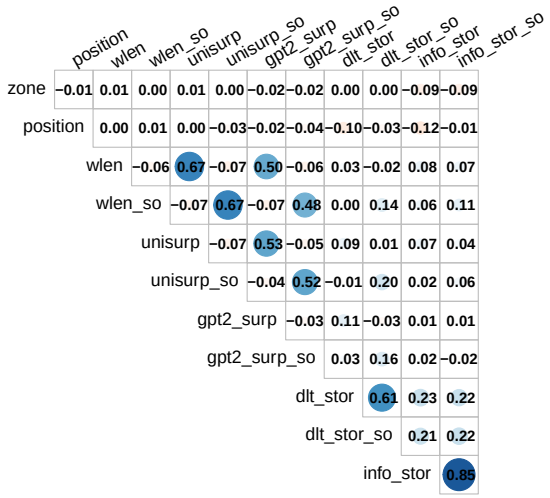
(3) a. The dancer who warned the guard questioned the president
b. The dancer who the guard warned questioned the president
(4) a. The senator who trusted the reporter visited the president
b. The senator who the reporter trusted visited the president
(5) a. The nurse who visited the thief stopped the reporter
b. The nurse who the thief visited stopped the reporter
(6) a. The singer who visited the actor warned the guard
b. The singer who the actor visited warned the guard
(7) a. The baker who watched the judge noticed the teacher
b. The baker who the judge watched noticed the teacher
(8) a. The reporter who helped the artist warned the singer
b. The reporter who the artist helped warned the singer
(9) a. The teacher who helped the lawyer noticed the banker
b. The teacher who the lawyer helped noticed the banker
(10) a. The doctor who helped the president attacked the chef
b. The doctor who the president helped attacked the chef
(11) a. The neighbor who met the thief noticed the teacher
b. The neighbor who the thief met noticed the teacher
(12) a. The dancer who warned the baker attacked the guard
b. The dancer who the baker warned attacked the guard
(13) a. The president who avoided the guard watched the soldier
b. The president who the guard avoided watched the soldier
(14) a. The thief who praised the neighbor visited the judge
b. The thief who the neighbor praised visited the judge
(15) a. The president who noticed the senator visited the writer
b. The president who the senator noticed visited the writer
(16) a. The soldier who attacked the president watched the student
b. The soldier who the president attacked watched the student
(17) a. The chef who watched the teacher avoided the baker
b. The chef who the teacher watched avoided the baker
(18) a. The neighbor who called the artist avoided the banker
b. The neighbor who the artist called avoided the banker
(19) a. The singer who questioned the writer attacked the actor
b. The singer who the writer questioned attacked the actor
(20) a. The actor who visited the artist praised the guard
b. The actor who the artist visited praised the guard
(21) a. The artist who avoided the baker called the chef
b. The artist who the baker avoided called the chef
(22) a. The singer who helped the guard avoided the student
b. The singer who the guard helped avoided the student
(23) a. The senator who called the student warned the detective
b. The senator who the student called warned the detective
(24) a. The president who praised the banker trusted the judge
b. The president who the banker praised trusted the judge
(25) a. The detective who warned the banker admired the actor
b. The detective who the banker warned admired the actor
(26) a. The actor who ignored the baker called the lawyer
b. The actor who the baker ignored called the lawyer
(27) a. The student who called the dancer stopped the guard
b. The student who the dancer called stopped the guard
(28) a. The dancer who questioned the judge trusted the pilot
b. The dancer who the judge questioned trusted the pilot
(29) a. The chef who stopped the neighbor ignored the student
b. The chef who the neighbor stopped ignored the student
(30) a. The driver who attacked the president watched the senator
b. The driver who the president attacked watched the senator

D Correlations between Predictors for Naturalistic Reading Analysis

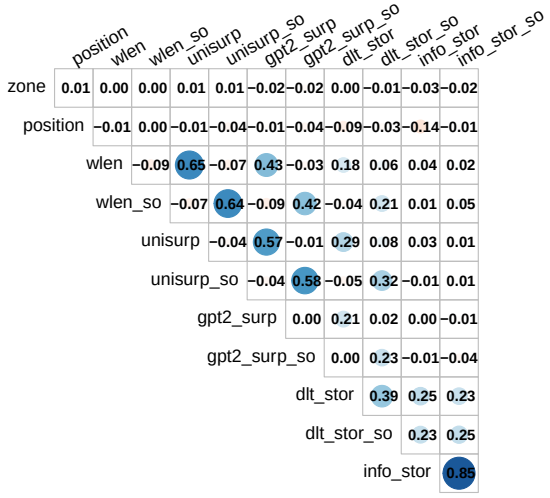
We present the Pearson correlation matrices for the predictors used in the analysis of the Natural Stories and OneStop datasets in Figure 7. The abbreviations for the predictors are defined as follows:

- `zone`: Word position index in the document.
- `position`: Word position index within the sentence.
- `wlen`: Word length in characters.
- `unisurp`: Unigram surprisal.
- `gpt2_surp`: Surprisal estimated by GPT-2 small.
- `dlt_stor`: Storage cost based on Dependency Locality Theory.

- `info_stor`: Information-theoretic storage cost (proposed).
- `_so`: The suffix appended to variables to denote their corresponding spillover terms.



(a) Natural Stories



(b) OneStop

Figure 7: Pearson correlation matrices of predictors in the naturalistic reading-time datasets.

E Comparison with other BERT-family models

While the main text reports the information storage estimates using the BERT-base model (`bert-base-uncased`), here we present a comparison with estimates derived from other models. Specifically, we report the information storage estimates by BERT-Large (`bert-large-uncased`) and RoBERTa (`roberta-base`; Liu et al., 2019). The architectural specifications of each model are summarized in Table 1.

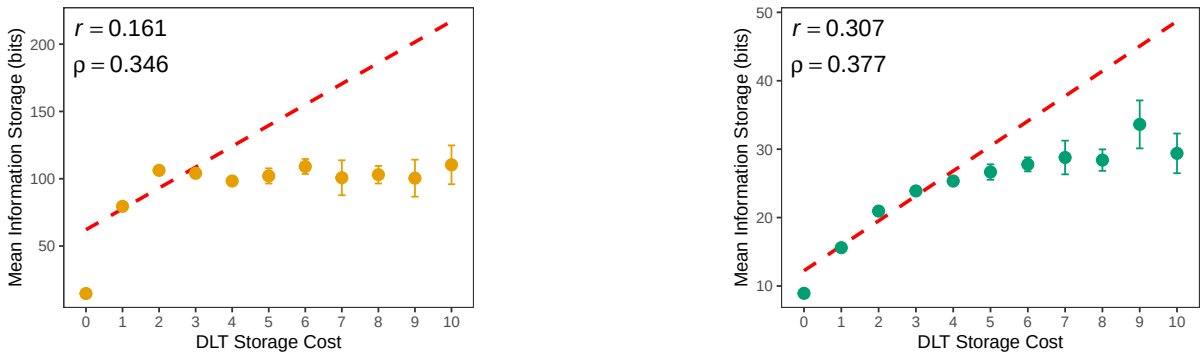
Model	Layers	Heads	Hidden Size	Parameters
BERT-base (`bert-base-uncased`)	12	12	768	110M
BERT-Large (`bert-large-uncased`)	24	16	1,024	336M
RoBERTa (`roberta-base`)	12	12	768	125M

Table 1: Architectural details of the models. Layers, Heads, and Hidden Size refer to the number of layers, the number of attention heads per layer, and embedding/hidden size, respectively.

E.1 Correlation to DLT storage

Following the methodology described in Section 4.2, we computed information storage on the UD_English-GUM corpus using BERT-Large and RoBERTa, and investigated its correlation with DLT storage cost. The results are illustrated in Figure 8.

For RoBERTa, although the correlation coefficients decreased slightly compared to the BERT-base results reported in Section 4.2, the overall trend remained largely consistent. In the case of BERT-Large, while the Pearson correlation coefficient was lower than those of the other models, the Spearman correlation remained above 0.3. This suggests that the relationship between information storage and DLT storage is not strictly linear, despite exhibiting a clear monotonically increasing trend.



(a) BERT-Large

(b) RoBERTa

Figure 8: Mean information-theoretic storage cost as a function of DLT storage cost in the UD_English-GUM corpus. Points represent the mean value for each DLT bin with 95% confidence intervals. For this visualization, bins with fewer than 100 observations were excluded due to data sparsity. The red dashed line represents the linear regression fitted to the raw data.

E.2 Reading-time analysis

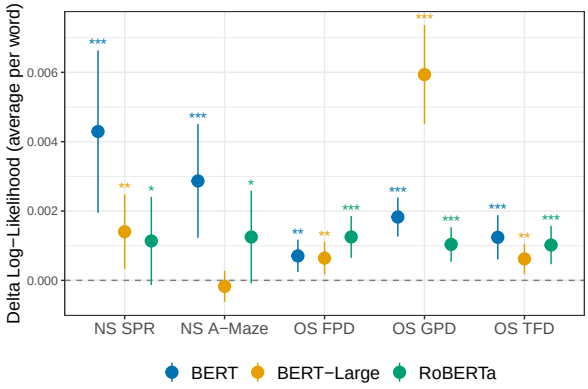
We replicated the reading-time analysis from Section 4.3 using the information storage estimates derived from the different BERT models. Specifically, we investigated whether information storage exhibits significant predictive power for reading times and examined the resulting coefficients. The results are presented in Figure 9.

The results for RoBERTa largely aligned with those for BERT-base. Although the magnitude of the effects varied, the Δ_{LL} values were all significantly greater than zero, indicating significant predictive power for reading time. Furthermore, the directions of the coefficients were entirely consistent, with the sole exception of the spillover coefficient in the Natural Stories SPR dataset.

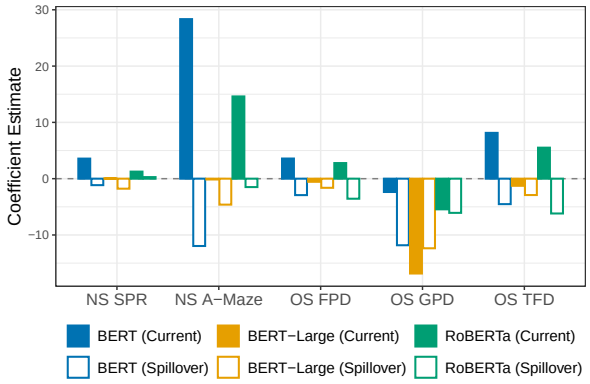
For BERT-Large, Δ_{LL} indicated significant predictive power for all datasets except Natural Stories A-Maze; however, the coefficients were predominantly negative. Given that DLT storage also exhibited a negative tend (see Figure 5b), the interpretation of these coefficients warrants further examination in future work.

Overall, BERT-base and RoBERTa demonstrated highly similar behavior. In contrast, while BERT-Large often patterned with BERT-base, their behavior was not always perfectly aligned. This discrepancy

can likely be attributed, at least in part, to the differences in parameter counts (see Table 1). This observation may be related to empirical findings on causal language models, which demonstrate that surprisal estimates from models with approximately 125M parameters provide the best fit for reading time in English, whereas larger models yield poorer fits (Oh and Schuler, 2023).



(a) Predictive power of information storage measures quantified by per-word Δ_{LL}. Information storage is estimated from three different BERT models: BERT (`bert-base-uncased`), BERT-Large (`bert-large-uncased`), and RoBERTa (`roberta-base`). Points and error bars represent means and 95% confidence intervals, respectively. Stars indicate FDR-adjusted significance levels (*** $q < .001$, ** $q < .01$, * $q < .05$, n.s. $q \geq 0.05$).



(b) Mean regression coefficients for information storage at the current and spillover regions, averaged across 10 CV folds. As the predictors were z-scored, coefficients represent relative effect sizes.

Figure 9: Results of the naturalistic reading-time analysis on Natural Stories (NS) and OneStop (OS). Abbreviations: SPR = self-paced reading; FPD = first-pass duration; GPD = go-past duration; TFD = total fixation duration.