

Peak Information Flux: A Resource Rational Measure of Language Processing Difficulty

Millan Hamilton

Gretchen Whitney High School
millan.parikh.hamilton@gmail.com

Jiaxuan Li

University of California, Irvine
jiaxul19@uci.edu

Abstract

A key goal in psycholinguistics is to quantify real-time language processing effort. A notable amount of literature links the processing effort to the amount of unexpected information carried by the new input. However, such theories fail to capture the dynamic evolution of comprehenders' interpretation which may involve revision and operate under cognitive constraints. We propose *peak information flux*, a metric that better captures the cost of reanalysis through a resource rational framework. We characterize language comprehension as continuous belief updating under finite information processing capacity and that the maximal rate of belief change signals language processing effort. We validate the performance of peak information flux on Syntactic Ambiguity Processing and Natural Stories corpora featuring self-paced reading time and a variety of eye-tracking metrics representing different stages of lexical processing. Peak information flux predicts reading behavior across multiple measures and complements surprisal in measures associated with reinterpretation and recovery. Additional analyses reveal that garden-path constructions exhibit delayed and asymmetric information flux dynamics, supporting the hypothesis of a sustained probabilistic revision process. Our model provides an information-theoretic metric for language comprehension under resource constraints.

 [Millan-H/peak-information-flux](#)

1 Introduction

Language comprehension is constrained by finite computational resources. Comprehenders continuously update probabilistic beliefs based on incomplete perceptual evidence and finite information processing capacity. A large body of work has argued that comprehension is often shallow or “good-enough” (Ferreira et al., 2002; Ferreira and Stacey,

2000). Comprehenders may rely on partial information to arrive at a more plausible alternative interpretation before constructing fully veridical interpretations (Ryskin et al., 2018, 2021; Futrell et al., 2020; Gibson et al., 2013; Levy, 2011). Such interpretations may be revised as additional evidence accumulates (Christianson et al., 2001; Norris and Kinoshita, 2008). These findings suggest that language comprehension is a dynamic process that involves continuous competition between multiple sources of information, uncertainty over perceptual input and cognitive constraints.

It has been challenging to quantify the complex dynamic process of language comprehension, especially when it involves reinterpretation. The probabilistic theories of language processing have linked the processing difficulty to the amount of new information conveyed by the input. In particular, surprisal theory proposes that language processing difficulty is linearly related to the unexpectedness of the current input given previous contextual information, often operationalized with contextualized next word probability in contemporary language models (Levy, 2008; Hale, 2001). The surprisal theory has successfully predicted language processing difficulties across multiple paradigms and languages (Frank and Bod, 2011; Smith and Levy, 2013; Boston et al., 2008; Wilcox et al., 2020; Shain et al., 2022; Wilcox et al., 2023; Xu et al., 2023; Michaelov and Bergen, 2020; Frank et al., 2013). However, surprisal-based theories characterize comprehension difficulty primarily in terms of static prediction error, and therefore do not fully capture the temporal dynamics of interpretive revision. Inputs with similar surprisal values can differ substantially in processing difficulty depending on the availability of competing interpretations and the extent of reinterpretation required to resolve ambiguity. This limitation is especially evident in garden-path constructions and recoverable anomalies, where difficulty often reflects prolonged com-

petition among interpretations rather than unexpectedness alone. [van Schijndel and Linzen \(2021\)](#) and [Huang et al. \(2024\)](#) have shown that surprisal tends to underestimate the size of garden path effects, and could not predict the magnitude of garden-path effects across different constructions.

A variety of theories have attempted to address language processing under uncertainty and cognitive pressure. [Futrell et al. \(2020\)](#) proposes that comprehenders maintain imperfect memory representations of prior context. Similarly, rational inference theories argue that comprehenders continuously evaluate multiple candidate interpretations under uncertainty, and that comprehension difficulty depends on the comparative probability of alternative interpretations ([Gibson et al., 2013](#); [Clark et al., 2025](#); [Ryskin et al., 2018](#); [Vigly et al., 2025](#)). [Giulianelli et al. \(2023, 2024\)](#) demonstrate that reading difficulty is sensitive to distances from plausible alternatives and that surprisal adjusted based on similarity to alternatives could account for reading times better. The Uniform Information Density hypothesis similarly links processing constraints to information rate, proposing that language users tend to distribute information relatively uniformly across linguistic units to avoid local peaks in processing demand ([Jaeger and Levy, 2006](#); [Meister et al., 2021](#)). Others have taken a resource rational approach to understand cognition under constraints ([Lieder and Griffiths, 2020](#); [Zaslavsky et al., 2018, 2021](#); [Cheyette and Piantadosi, 2020](#); [Sims, 2016, 2018](#)). Memory-based approaches have also been tried, such as embedding depth in parser models ([van Schijndel et al., 2012](#)), modeling constraints on total memory storage. In particular, [Li and Futrell \(2024\)](#) formalize comprehension as an optimal tradeoff between interpretation accuracy and processing depth, and link processing effort to the rate of change in processing depth.

In this paper, we propose an information-theoretic metric *peak information flux*, or PIF, that quantifies language processing effort under cognitive constraints. We model comprehension as the evolution of an interpretation distribution over possible meanings as evidence accumulates over time. The processing depth is constrained by a finite information-processing capacity, which in turn constrains the rate at which processing depth can increase. We define the rate of change in processing depth as the information processing flux, representing the rate at which the comprehender reorganizes probabilistic beliefs during interpretation.

We further propose that comprehension effort can be approximated by the peak information flux: the maximal instantaneous processing demand encountered during processing. We test this hypothesis on two large-scale reading datasets and compare peak information flux against surprisal as a predictor of reading behavior across multiple reading-time measures. Our results show that peak information flux consistently predicts human reading times and captures variance beyond standard surprisal-based accounts. In syntactic ambiguity constructions, the model explains differences in garden-path difficulty as a consequence of sustained probabilistic reinterpretation rather than the total amount of new information (surprisal) alone. Our findings suggest that language comprehension difficulty is shaped by the temporal dynamics of belief updating under finite processing bandwidth.

2 Model

Belief update We hypothesize that language comprehension involves instantaneous and gradual belief updating upon receiving perceptual evidence. We formalize this process as updating a probability distribution over candidate interpretations. Let c denote the preceding linguistic context, x the observed word, and $w \in \mathcal{W}$ a possible candidate interpretation of that word. Before processing x , the comprehender has a prior distribution $p_0(w | c)$. During processing, perceptual evidence from x modifies this distribution. We denote the interpretation distribution after t units of model processing by $p_t(w | x, c)$.

Processing depth We assume that changing an interpretation away from prior expectations requires information processing. Following a rate-distortion view ([Zaslavsky et al., 2021](#); [Lieder and Griffiths, 2020](#); [Vigly et al., 2025](#)), we quantify the amount of such processing by the information gained in moving from the prior distribution p_0 to the current interpretation distribution p_t . For an input x , we define its processing depth at time t as

$$D(t) = D_{\text{KL}}[p_t(w | x, c) \parallel p_0(w | c)]. \quad (1)$$

Importantly, we assume that this belief revision cannot occur without limit. At any point during processing, the comprehender has access to a finite amount of information-processing capacity $C(t)$. The interpretation distribution is therefore

The storyteller could turn any incident into an amusing ...

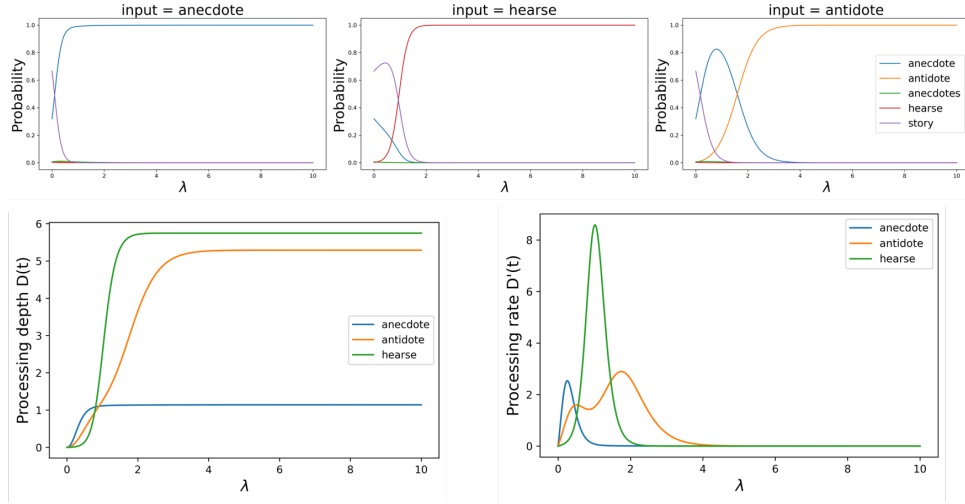


Figure 1: Illustration of model behavior for three continuations in the context *The storyteller could turn any incident into an amusing ...*. The top panels show how the interpretation policy $p(w|x)$ evolves as processing depth increases with λ . The bottom panels show cumulative processing depth $D(t)$ and processing rate $D'(t)$. The visualization is adapted from Li and Futrell (2024) but using a differing candidate set and including additional visualizations. The example sentence is adapted from Ryskin et al. (2021).

constrained such that:

$$D(t) \leq C(t). \quad (2)$$

The constraint does not imply that comprehenders deliberately calculate KL divergence. Rather, KL divergence provides an information-theoretic description of the extent to which perceptual evidence has been allowed to alter prior expectations.

Optimal interpretation policy The goal of language comprehension is to minimize the distortion between internal interpretation and perceptual input. Let $d(w, x)$ denote the distortion between candidate interpretation w and perceptual input x .

At a given processing depth, comprehension minimizes expected distortion given the constraint of information-processing capacity:

$$\min_{p_t(w|x)} \mathbb{E}_{p(x)p_t(w|x)} [d(w, x)] \text{ s.t. } D(t) \leq C(t). \quad (3)$$

The constrained optimization has the solution:

$$p_t(w | x) \propto p_0(w) e^{-\lambda(t) d(w, x)}, \quad (4)$$

where $\lambda(t)$ controls the relative influence of perceptual evidence and prior expectations. This formulation is closely related to the rate–distortion account of language comprehension developed by Li and Futrell (2023, 2024).

The optimal interpretation policy offers an intuitive interpretation of processing depth. When

$\lambda = 0$, the interpretation is determined entirely by prior expectations and $D(t) = 0$. As λ increases, candidates with greater distortion are progressively penalized and perceptual evidence exerts more influence. Because the observed word itself has zero distortion from x , sufficiently deep processing concentrates probability on the veridical input. Thus, increasing λ will shift comprehenders’ interpretation from a shallow, expectation-driven one to a veridical one.

Information flux and processing effort The temporal dynamics defined above are central to account for the processing effort. Consider two inputs that ultimately require the same total departure from prior expectations. For one input, the change might occur gradually over the processing trajectory. For another, a large amount of probability mass might need to be redistributed over a short interval. A resource-limited processing system should be more strongly challenged in the second case even when the final processing depth is comparable.

We define information flux as the instantaneous rate of change in processing depth $\frac{dD(t)}{dt}$. Information flux measures how rapidly information from the perceptual input is changing the comprehender’s interpretation. We argue that processing effort should be particularly sensitive to the largest instantaneous demand placed on a finite-capacity

processing system. We propose that the processing effort could be quantified as the *peak information flux*:

$$\max_t F(t) = \max_t \frac{dD(t)}{dt}. \quad (5)$$

Peak information flux and surprisal characterize different aspects of processing. Surprisal quantifies the amount of information provided by the new input that actually occurs. Peak information flux instead depends on the distribution of probability over competing interpretations, their similarity to the observed input, and the trajectory through which that distribution changes as processing deepens. Consequently, two equally surprising words can have different peak information flux. If the alternatives favored by context are readily distinguishable from the input, probability can move directly toward the target. If a highly expected alternative closely resembles the target, the system may initially maintain the expected interpretation and only later revise it, producing a qualitatively different information-flux trajectory.

Figure 1 illustrates this distinction. Following *The storyteller could turn any incident into an amusing...*, the expected continuation *anecdote* is strongly supported by the context. For the form-similar but unexpected input *antidote*, early processing can therefore favor *anecdote* before evidence accumulates in favor of *antidote*. The resulting processing trajectory contains an initial movement toward the plausible alternative followed by subsequent reinterpretation of the input. By contrast, an unexpected continuation whose perceptual evidence strongly distinguishes it from the expected alternatives can produce a more direct redistribution of probability. Peak information flux is designed to quantify these differences in the dynamics of reinterpretation. We hypothesize that processing difficulty increases when comprehension requires a large amount of probabilistic belief revision to occur within a narrow portion of the processing trajectory.

Implementation The framework above requires three quantities for each observed word: a set of candidate interpretations $\mathcal{W}$, a prior probability $p_0(w \mid c)$ over those candidates, and a distortion function $d(w, x)$.

We approximate the set of interpretations that may compete during word recognition using three complementary sources. For each target word, we

retrieve: the 100 highest-probability contextual continuations from GPT-2 small given the preceding context (Radford et al.); 100 semantically similar words based on GloVe cosine similarity (Pennington et al., 2014); and 100 form-similar words selected from the SUBTLEX-US dataset (Brysbaert and New, 2009) using character-level Levenshtein edit distance. The observed target word is also explicitly included. We then combine and deduplicate these sets, producing at most 301 candidates for a target position. We use GPT-2 small to estimate contextual expectations. These probabilities are renormalized over the candidate set. Distortion is operationalized using character-level Levenshtein edit distance (Bachmann, 2025).

We parameterize the processing trajectory by increasing $\lambda(t) = \lambda_0 t$, where $\lambda_0 = \lambda_{\max}/T$, T is the number of processing steps, and $\lambda_{\max}$ is chosen sufficiently large for the interpretation distribution to converge. We evaluate $D(t)$ along this trajectory and obtain the corresponding information-flux trajectory from changes in processing depth across successive steps. For each dataset, we assign a single T that is large enough for the interpretation distributions to converge.

3 Study 1: Eye-Tracking on Syntactic Ambiguity Processing

3.1 Dataset

We use the Syntactic Ambiguity Processing Benchmark (Timkey et al., 2026) collected from 424 subjects featuring different types of syntactically ambiguous sentences compared with counterpart control. It collects word-by-word data via eye trackers on the time spent reading each word, collecting 4 measures: the forward reading time for word n , a binary measure encoding whether that word is a regression, the total fixation time on word n where word n is a word regressed from, and the duration of all first fixations on word n and the duration of fixations on words regressing from word n . We use these data to benchmark our model against surprisal in predicting reanalysis difficulty.

3.2 Experimental Setup

We analyzed five eye-tracking measures: forward reading time, gaze duration, go-past time, first-pass regressions out, and total reading time. Prior to analysis, we removed observations by metric by construction greater than 3 standard deviations from that metric’s mean, excluded the first word

of each passage, discarded missing reading-time observations, and excluded data from participants whose comprehension-question accuracy fell below 80%. For regressions out, we used total reading time as the metric for filtering by standard deviation since regressions out isn't time based.

We fit linear mixed-effects regression models (Bates et al., 2015). The baseline model included the log frequency of current and previous word derived from SUBTLEX-US using Lg10WF as well as the current word position. We then compared this baseline against models additionally including either peak information flux or surprisal as predictors. All models are maximally converging with by-subject and by-item random effects. For each model, we report BIC values to compare their performance using REML = FALSE. P-values for fixed-effect slopes were calculated via lmerTest, which implements Satterthwaite's approximation (Kuznetsova et al., 2017). Optimization made use of the bobyqa optimizer with a maximum of 200,000 function evaluations.

We use this dataset with peak information flux. We evaluated against surprisal, calculated as standard $-\log p(w \mid \text{context})$ using GPT-2. For multi-token words, we summed the surprisals of each token. We fit each predictor against forward reading time, gaze duration, go-past time, first-pass regressions out, and total reading times from the Syntactic Ambiguity Processing Benchmark. In addition to the aforementioned predictors, we also fit $\max_t D'(t) + \text{surprisal} + \text{controls}$ to each metric in order to test whether $\max_t D'(t)$ explains additional variance in regression times beyond surprisal. We do these evaluations by metric constrained to just the critical and spillover words fitting one combined model on both word types for each metric due to sample size constraints for either metric alone. We also do evaluations by three types of garden path sentences: Main Verb/Reduced Relative (MV/RR), Intransitive/Transitive (NP/Z) and Noun Phrase/Sentential Complement (NP/S).

3.3 Eye-Tracking Regression Results

Table 1 compares surprisal and peak information flux (PIF) as predictors of reading behavior across five eye-tracking measures and three garden-path constructions. Overall, PIF yields lower BIC in 10 of the 15 comparisons. The clearest advantages for PIF occur for MV/RR total reading time ($\Delta \text{BIC} = 9.97$), NP/S go-past time (4.29), and MV/RR forward reading time (3.40). In contrast, surprisal

Construction	N	BIC		Δ BIC
		Surprisal	$\max_t D'(t)$	
<i>Forward reading time</i>				
MV/RR	985	12657.24	12653.84	3.40
NP/S	1,353	17078.50	17079.33	−0.83
NP/Z	1,060	13450.22	13457.71	−7.49
<i>Gaze duration</i>				
MV/RR	1,874	24074.75	24073.60	1.15
NP/S	1,894	23981.08	23980.83	0.25
NP/Z	1,864	23699.56	23709.50	−9.94
<i>Go-past time</i>				
MV/RR	1,865	30585.78	30585.77	0.01
NP/S	1,886	28639.47	28635.18	4.29
NP/Z	1,858	30388.13	30387.29	0.84
<i>First-pass regressions out</i>				
MV/RR	1,864	2588.59	2587.54	1.05
NP/S	1,897	2266.88	2265.83	1.06
NP/Z	1,860	2556.40	2556.50	−0.10
<i>Total reading time</i>				
MV/RR	2,014	31639.16	31629.19	9.97
NP/S	2,067	30608.08	30610.14	−2.06
NP/Z	2,046	31389.62	31389.47	0.16

Table 1: BIC for the surprisal and $\max_t D'(t)$ models, fit separately per garden-path construction (MV/RR, NP/S, NP/Z) on the full critical + spillover region, with the better-fitting model in bold. ΔBIC is computed as $\text{surprisal} - \max_t D'(t)$, so positive values favor $\max_t D'(t)$. N represents the total amount of observations on which each model was fit. The ΔBIC values were calculated with greater precision, meaning they may not be the perfect difference between the other two columns.

is strongly favored for NP/Z gaze duration (-9.94) and forward reading time (-7.49). Differences for first-pass regressions are negligible across all three constructions. Joint models containing both predictors further indicate that their contributions depend on construction and reading measure. PIF explains unique variance in NP/S go-past time and MV/RR total reading time, whereas surprisal explains unique variance in NP/Z forward reading time and gaze duration (see Appendix A). These results suggest that PIF and surprisal capture partially overlapping but dissociable components of processing difficulty, with PIF providing additional explanatory value in some conditions associated with later processing and reinterpretation.

3.4 Analysis of Garden Path Effect

To capture the model performance on the garden path effect, we use the generalization paradigm developed in Timkey et al. (2026). First, we fit linear mixed-effects models on the filler sentences from the Syntactic Ambiguity Processing Bench-

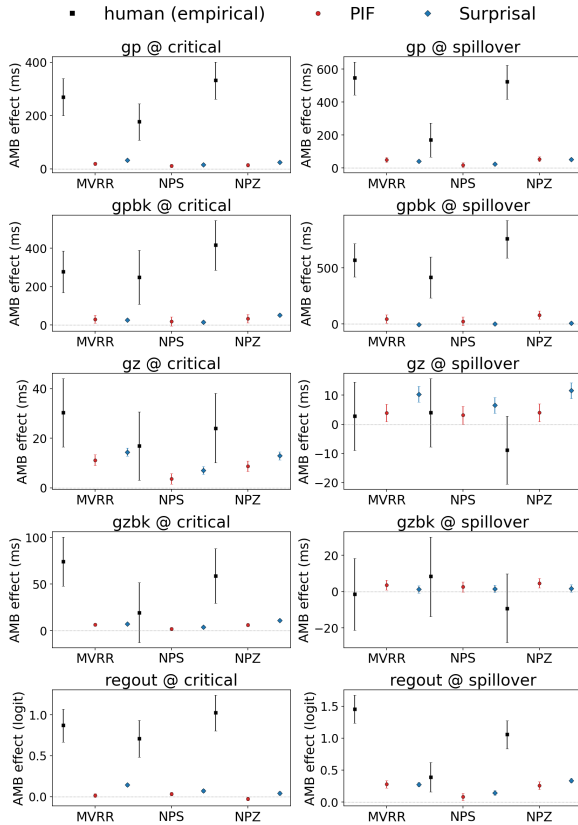


Figure 2: Empirical and simulated garden path effect size derived from the generalization paradigm. "regout" represents regressions out, "gz" and "gzbk" represent gaze duration and regressive gaze duration, a variant restricted to only words where a regressive saccade occurred, and "gp" and "gpbk" are go-past and regressive go-past times. "Human" refers to the empirical garden path effect computed from SAP's reading time data.

mark to estimate baseline reading-time effects for different predictors. We separately fit models using surprisal and peak information flux as predictors, as well as a joint model containing both predictors. Second, we analyzed the garden-path sentences using Bayesian models through the brms R package (Bürkner, 2021). For each word in the ambiguous and unambiguous sentence versions, the dependent variable was the reading-time prediction generated by the corresponding linear mixed-effects model. The primary predictor was sentence ambiguity. This procedure allowed us to evaluate the extent to which each predictor captures the reading-time differences associated with syntactic ambiguity.

Fig. 2 compares the empirical garden-path effect with effects predicted from surprisal and PIF. Both predictors substantially underestimate the empirical effect across several measures. PIF is numeri-

cally closer to the human effect for some regressive measures at the spillover region, particularly go-past and regressive go-past time. However, confidence intervals overlap broadly, and neither predictor shows a consistent advantage across measures. Thus, PIF shows limited improvement in capturing the magnitude of garden-path effects, particularly in some later processing measures, but does not fully account for the observed disambiguation cost.

4 Study 2: Self-Paced Reading on Natural Stories

In Study 2, we extend the analysis to the Natural Stories Corpus, a large-scale naturalistic self-paced reading dataset containing many rare and syntactically complex constructions (Futrell et al., 2018).

4.1 Dataset

We evaluated on the Natural Stories dataset, containing 10 stories of naturalistic text and self-paced reading data from 181 subjects. The stories contain a total of 10,256 words. There are only a total of 800,480 observations due to our filtering process and the fact that not all participants read all stories in the dataset, with an initial N of 848,875.

4.2 Experimental Setup

The setup of linear mixed-effects models is similar to Study 1. We also include additional controls, with everything in the Eye-Tracking models as well as word length and previous word length. We also included a more defined random effects structure, with random intercepts and slopes without correlations for both subjects and items.

4.3 Results

Predictor	N	BIC
surprisal	800,480	9480188.16
$\max_t D'(t)$	800,480	9479998.00
ΔBIC		190.16

Table 2: BIC for surprisal and information flux on Natural Stories self-paced reading. N represents the total amount of observations on which each model was fit. ΔBIC shows surprisal $-\max_t D'(t)$.

Table 2 displays our ΔBIC values on the Natural Stories dataset for self-paced reading time. On Natural Stories, the standalone PIF model yields a lower BIC than the surprisal model ($\Delta BIC =$

D'' magnitude asymmetry ($\max D'' + \min D''$) by difficulty quintile

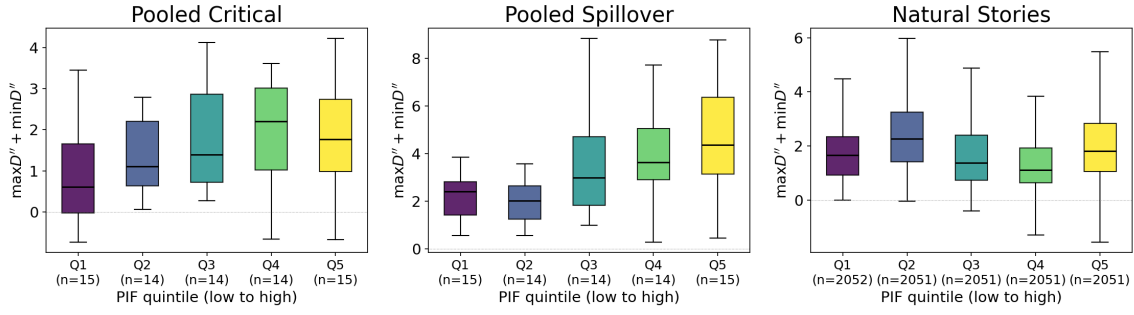


Figure 3: D'' magnitude asymmetry, $\max D''(t) + \min D''(t)$, binned by PIF quintile for the three datasets: pooled critical, pooled spillover, and Natural Stories. The metric is zero for words whose $D''(t)$ peak and trough are symmetric in magnitude. Positive values indicate that the rise in $D'(t)$ is sharper than its fall, and negative values indicate a fall in $D'(t)$ that's sharper than its rise. Pooled means this includes data from the specified region of all constructions, MV/RR, NP/S, and NP/Z, in the figure.

Distribution of time of PIF: λ at which PIF occurs

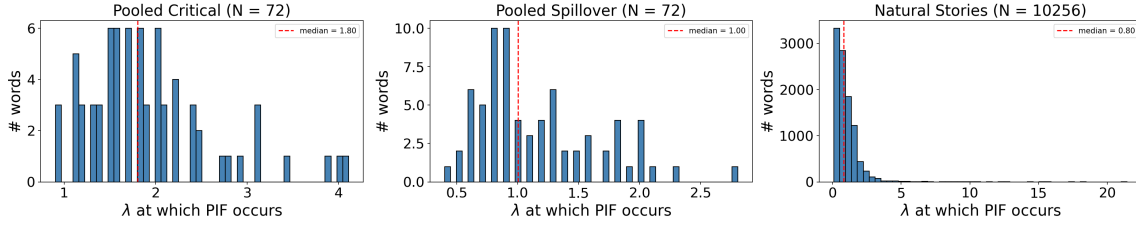


Figure 4: Distribution of $\arg \max_t D'(t)$ for the three datasets. This shows the time of information flux across garden path critical, garden path spillover, and naturalistic texts. Similar to magnitude asymmetry, pooled means the plot uses data from the specified region of all three constructions.

190.16). In the joint model, however, surprisal remains a significant predictor (partial $r = .723$), whereas PIF does not ($r = -.078$; Appendix B). PIF predicts naturalistic reading times well when considered independently, but does not explain substantial unique variance beyond surprisal in this dataset.

5 Information Flux Analysis

We have established peak information flux as a predictor of reading-time measures, particularly in metrics associated with reinterpretation and recovery. Here we examine the temporal dynamics surrounding information flux in order to better understand why larger information flux values are associated with greater processing difficulty. We analyze two properties of the information flux timecourse: Magnitude asymmetry and the timing of information flux.

Magnitude asymmetry. To investigate the recovery dynamics of belief updating, we analyze magnitude asymmetry, defined as $\max_t D''(t) +$

$\min_t D''(t)$. It measures the relative concentration of increase and decrease of information flux $D'(t)$ over time. Positive magnitude asymmetry suggests the rise of information flux is steeper than its fall.

The magnitude asymmetry characterizes how rapidly processing demand accelerates and decays over the course of interpretation. Under our framework, processing concludes when information flux gradually approaches zero. Consequently, if the fall of information flux is slower than its rise, processing should remain active for a longer duration, extending reading times. We therefore examine whether more difficult sentence types exhibit greater asymmetry, reflecting prolonged residual processing after rapid belief revision.

Time of information flux. We also analyze the time of peak information flux in order to determine whether difficult inputs exhibit delayed onset of maximal belief updating. The time of information flux is defined as $\arg \max_t D'(t)$, which represents the λ in the timecourse at which peak information flux is reached. Under our framework, later peak

information flux may reflect the additional processing required before substantial reinterpretation can occur, such as competition among interpretations or delayed integration of incoming evidence. This analysis allows us to examine when reading difficulty occurs during the timecourse of processing.

5.1 Results

Figure 3 shows that magnitude asymmetry generally increases across PIF quintiles for garden-path items, particularly in the spillover region, whereas the relationship is substantially weaker in Natural Stories. Figure 4 likewise shows later peak information flux for garden-path critical words (median $\lambda = 1.80$) than for spillover words (1.00) or Natural Stories (0.80). These analyses characterize the trajectories associated with high information flux: garden-path items tend to exhibit rapid increases followed by slower decreases in processing rate and later peaks in belief revision.

5.2 Interpretation

The temporal analyses provide a descriptive characterization of the processing trajectories associated with peak information flux. Garden-path items, particularly in the spillover region, tend to show greater magnitude asymmetry than words in Natural Stories: processing demand rises relatively sharply and decays more gradually. Garden-path critical words also reach peak information flux later in the model trajectory. Together, these patterns are consistent with the idea that difficult inputs involve more prolonged and temporally distributed belief revision rather than a single instantaneous update.

The construction-level analyses show a similar pattern. MV/RR and NP/Z constructions, which exhibit larger garden-path effects in the behavioral data, also tend to show greater asymmetry and later peak information flux than NP/S constructions (Appendix E). This correspondence suggests that the temporal structure of belief updating may help characterize differences in reanalysis difficulty across constructions.

These analyses should, however, be interpreted descriptively. Magnitude asymmetry, peak timing, and peak information flux are derived from the same underlying processing-depth trajectory and therefore do not constitute independent evidence that delayed or asymmetric updating causes longer reading times. Rather, they clarify what distinguishes high-information-flux trajectories and motivate future work directly testing whether the

timing and duration of belief revision explain behavioral variation beyond peak magnitude alone.

6 Discussion

We introduce peak information flux, an information-theoretic measure of language processing effort under finite processing bandwidth. Concretely, we characterize language comprehension as an optimal trade-off between processing depth and accuracy, and that language processing effort could be indexed by the maximal rate of change in comprehenders' processing depth. We validate this hypothesis on two large-scale reading datasets, featuring eye movements over syntactically ambiguous garden path sentences and self-paced reading times to naturalistic text with rare constructions. Peak information flux predicts various human signatures of reading difficulties, and complements surprisal in accounting for later stage of language processing. We further analyzed the asymmetry of the information flux and the timepoint when peak information flux is reached, which suggest that syntactically ambiguous sentences involve delayed language reinterpretation.

Our measure differs from surprisal accounts, which characterize difficulty primarily in terms of the unexpectedness of the observed input. The key distinction is between information content and information-processing rate. Surprisal quantifies how much information is conveyed by an input relative to prior expectations. PIF instead characterizes the temporal dynamics of the belief update required to incorporate that input. Two words can therefore have similar surprisal but differ in PIF if they induce different trajectories through competing interpretations. Under this view, processing difficulty reflects not only prediction error, but the dynamics of resolving competition among interpretations.

This proposal places PIF within a broader class of theories that treat language comprehension as computation under limited resources, while locating the resource constraint at a particular level. Uniform Information Density (UID) proposes that linguistic systems tend to avoid large fluctuations in information rate across a sequence, reflecting limits on how much information can be transmitted or processed at once (Jaeger and Levy, 2006; Meister et al., 2021). PIF shares the basic intuition that concentrated information demand is costly, but shifts

to the distribution of belief revision within the processing of an input. In this sense, UID concerns the temporal organization of information in the signal, whereas PIF concerns the temporal organization of inference in the comprehender. The moderate correlations between our UID operationalization and PIF in SAP ($r = .239$) and in Natural Stories ($r = .396$) are consistent with the two measures capturing related but non-identical aspects of information concentration (Appendix D).

Our model argues that language comprehension operates under a finite processing capacity and that comprehension difficulty emerges when the rate of required belief updating approaches finite processing capacity. The idea of bounded rationality has been widely applied to understand heuristics and cognitive bias in psychology (Sims, 2016, 2018; Zaslavsky et al., 2018, 2021; Mollica et al., 2021; Cheyette and Piantadosi, 2020; Bhui et al., 2021; Arumugam et al., 2022; Futrell, 2023). In the domain of language, it has been argued that language systems efficiently compress ideas into words by optimizing the information bottleneck (IB) trade-off between the complexity and accuracy of the lexicon (Kemp and Regier, 2012; Regier et al., 2007; Zaslavsky et al., 2018; Zipf, 1949; Piantadosi, 2014; Mahowald et al., 2013). The accuracy-effort trade-off has been used to explain mental representation of languages. Similar information-theoretic principles have also been applied in the domain of language (Kemp and Regier, 2012; Regier et al., 2007; Zaslavsky et al., 2018; Zipf, 1949; Piantadosi, 2014; Mahowald et al., 2013).

Most relevant to our work, Li and Futrell (2024) proposes a resource rational framework of real-time language comprehension, where comprehenders find optimal interpretations of their input given a constraint on how much information they process from the input. The present proposal adds a dynamic hypothesis to this framework. PIF demonstrates that the same eventual interpretation can be reached through a gradual trajectory or through a rapid redistribution of probability mass. We hypothesize that these trajectories differ in processing cost because a finite-capacity system is particularly challenged when a large amount of updating must occur at once.

The eye-tracking results provide partial support for this distinction. PIF and surprisal share substantial variance, but their predictive contributions are not identical. PIF shows selective advantages in some conditions associated with later processing,

including NP/S go-past time and MV/RR total reading time, whereas surprisal performs better in some earlier measures, particularly for NP/Z constructions. This pattern is consistent with a distinction between prediction error and interpretive revision. Surprisal is sensitive to the probability of the observed word, whereas PIF is sensitive to how probability mass over alternative interpretations changes after the input begins to constrain interpretation.

The temporal analyses further clarify what high-PIF processing trajectories look like. Garden-path items tend to exhibit more asymmetric information-flux trajectories and later peaks than words in Natural Stories. The model represents it as a temporally extended process in which competing interpretations remain active and probability is progressively reorganized.

This perspective also suggests a particular account of reanalysis. Garden-path difficulty may arise when an initially favored interpretation becomes increasingly incompatible with incoming evidence, forcing the system to redirect probability toward an alternative. PIF captures the intensity of this redistribution, but the current implementation represents competition primarily among lexical candidates rather than explicit syntactic structures. This limitation is important because both PIF and surprisal continue to underestimate the empirical magnitude of garden-path effects. A stronger model of syntactic reanalysis would represent competing structural hypotheses directly and track information flow between them. Under such a model, the cost of a garden path could arise not simply from encountering an unexpected word, but from rapidly reallocating probability between incompatible structural analyses.

Limitations

Our model still underestimates the empirical garden path effect size. One possibility is that peak information flux captures the redistribution of probability mass among competing interpretations, but not the additional cost of reconstructing initially discarded syntactic representation. Alternatively, this could be related to the current candidate generation process. We construct alternatives based on semantically or phonologically related words, and the current candidate generation procedure does not explicitly represent alternative syntactic analyses. This may have limited the model's ability to capture the full magnitude of garden path

effects. Future work should construct syntactic parses or sentence-level interpretations, which allows the model to track probability shift between competing structural interpretations. In addition, different candidate sets may provide different results, so future work should cover secondary and ablation analyses about how reduced, expanded, or differently composed candidate sets influence the performance of information flux. Specifically, expanded candidate sets should be explored, since our current work is limited to a relatively constrained set of a maximum of 301 candidates due to computational constraints. Beyond our model’s specifics, our current work lacks multilingual testing and multi-model testing. Our corpora are English-only, so future work should include expansion into other languages, ideally with different typological structures from English. Our analyses are constrained to GPT-2 as a result of compute constraints. Further work should include testing on a broader selection of models and architectures, such as recurrent neural networks. We chose to use GPT-2 since prior work substantiates that it is optimal for modeling human linguistic processing (Oh and Schuler, 2023), though this does not mean multi-model testing lacks value. Regarding implementation, due to compute and memory constraints, we also use a context window of 256 tokens with a 32 token stride rather than GPT-2’s full 1024 tokens. This may introduce problems for the Natural Stories dataset, where stories may exceed this window. Our current results in section 5 are also purely descriptive since they’re derived from the same trajectory as PIF, meaning the analyses are not independent. Further work should be done to investigate whether the temporal dynamics surrounding PIF are useful predictors of reading time.

Acknowledgements

We thank the reviewers for providing insightful feedback during the review process and helping us improve our paper. We also thank Richard Futrell for helpful discussions. AI models were used for assistance with implementation, to suggest candidate papers for related work, and for proofreading and revision help. Analysis of the literature and results was done by the authors, and all ideas in the paper come from the authors. Outputs were verified by the authors for correctness.

References

- Dilip Arumugam, Mark K. Ho, Noah D. Goodman, and Benjamin Van Roy. 2022. [On rate-distortion theory in capacity-limited cognition reinforcement learning](#). *Preprint*, arXiv:2210.16877.
- Max Bachmann. 2025. [rapidfuzz/rapidfuzz: Release 3.13.0](#).
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. [Fitting linear mixed-effects models using lme4](#). *Journal of Statistical Software*, 67(1):1–48.
- Rahul Bhui, Lucy Lai, and Samuel J Gershman. 2021. [Resource-rational decision making](#). *Current Opinion in Behavioral Sciences*, 41:15–21. Value based decision-making.
- Marisa Ferrara Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1).
- Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.
- Paul-Christian Bürkner. 2021. [Bayesian item response modeling in R with brms and Stan](#). *Journal of Statistical Software*, 100(5):1–54.
- Samuel J Cheyette and Steven T Piantadosi. 2020. A unified account of numerosity perception. *Nature Human Behaviour*, 4(12):1265–1272.
- Kiel Christianson, Andrew Hollingworth, John F. Halliwell, and Fernanda Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42(4):368–407.
- Thomas Hikaru Clark, Jacob Hoover Vigly, Edward Gibson, and Roger Levy. 2025. Resource-rational noisy-channel language processing: Testing the effect of algorithmic constraints on inferences. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23659–23672.
- Fernanda Ferreira, Karl GD Bailey, and Vittoria Ferraro. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1):11–15.
- Fernanda Ferreira and Janis Stacey. 2000. The misinterpretation of passive sentences. Manuscript currently unpublished.
- Stefan Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

- Stefan Frank, LJ Otten, G Galli, and G Vigliocco. 2013. Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 878–883.
- Richard Futrell. 2023. Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, 120(39):e2220593120.
- Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. 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 Piantadosi, and Evelina Fedorenko. 2018. [The natural stories corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Edward Gibson, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.
- Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024. [Generalized measures of anticipation and responsivity in online language processing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11648–11669, Miami, Florida, USA. Association for Computational Linguistics.
- Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. [Information value: Measuring utterance predictability as distance from plausible alternatives](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.
- 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*.
- Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. [Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty](#). *Journal of Memory and Language*, 137:104510.
- T. Jaeger and Roger Levy. 2006. [Speakers optimize information density through syntactic reduction](#). In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.
- Charles Kemp and Terry Regier. 2012. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054.
- Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. [lmerTest package: Tests in linear mixed effects models](#). *Journal of Statistical Software*, 82(13):1–26.
- Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
- Roger Levy. 2011. Integrating surprisal and uncertainty input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065.
- Jiaxuan Li and Richard Futrell. 2023. A decomposition of surprisal tracks the N400 and P600 brain potentials. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.
- Jiaxuan Li and Richard Futrell. 2024. [An information-theoretic model of shallow and deep language comprehension](#). In *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*.
- Falk Lieder and Thomas L Griffiths. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1.
- Kyle Mahowald, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.
- Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. [Revisiting the Uniform Information Density hypothesis](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- James Michaelov and Benjamin Bergen. 2020. How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663.
- Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp. 2021. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, 118(49):e2025993118.
- Dennis Norris and Sachiko Kinoshita. 2008. Perception as evidence accumulation and bayesian inference: insights from masked priming. *Journal of Experimental Psychology: General*, 137(3):434.
- 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.

- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Steven T Piantadosi. 2014. Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21:1112–1130.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*.
- Terry Regier, Paul Kay, and Naveen Khetarpal. 2007. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4):1436–1441.
- Rachel Ryskin, Richard Futrell, Swathi Kiran, and Edward Gibson. 2018. Comprehenders model the nature of noise in the environment. *Cognition*, 181:141–150.
- Rachel Ryskin, Laura Stearns, Leon Bergen, Marianna Eddy, Evelina Fedorenko, and Edward Gibson. 2021. An erp index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, 158:107855.
- Cory Shain, Idan A Blank, Evelina Fedorenko, Edward Gibson, and William Schuler. 2022. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *Journal of Neuroscience*, 42(39):7412–7430.
- Chris R Sims. 2016. Rate–distortion theory and human perception. *Cognition*, 152:181–198.
- Chris R Sims. 2018. Efficient coding explains the universal law of generalization in human perception. *Science*, 360(6389):652–656.
- Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.
- William Timkey, Kuan-Jung Huang, Byung-Doh Oh, Grusha Prasad, Suhas Arehalli, Tal Linzen, and Brian Dillon. 2026. [Eye movements reveal a dissociation between prediction and structural processing difficulty in language comprehension](#). *Proceedings of the National Academy of Sciences*, 123(32):e2532230123.
- Marten van Schijndel, Andy Exley, and William Schuler. 2012. [Connectionist-inspired incremental PCFG parsing](#). In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 51–60, Montréal, Canada. Association for Computational Linguistics.
- Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.
- Jacob Vigly, Peng Qian, Morgan Sonderegger, and Timothy O’Donnell. 2025. Comprehension effort as the cost of inference.
- Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. [On the predictive power of neural language models for human real-time comprehension behavior](#). In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020*. cognitivesciencesociety.org.
- Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. [Testing the predictions of surprisal theory in 11 languages](#). *Transactions of the Association for Computational Linguistics*, 11:1451–1470.
- Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.
- Noga Zaslavsky, Jennifer Hu, and Roger P. Levy. 2021. [A Rate–Distortion view of human pragmatic reasoning?](#) In *Proceedings of the Society for Computation in Linguistics 2021*, pages 347–348, Online. Association for Computational Linguistics.
- Noga Zaslavsky, Charles Kemp, Terry Regier, and Naf-tali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.
- George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

A Study 1: Additional analysis

Table 3 reports partial r values and p values for each predictor across the metrics we report. This is to analyze, in a joint model, which metric carries more information when sharing variance.

Across many of these metrics, neither predictor shows significance. However, on both regressive measures, $\max_t D'(t)$ appears to subsume surprisal on NP/S, with NP/Z on go-past also appearing to show this effect. On gaze duration metrics, surprisal appears to almost subsume $\max_t D'(t)$ on NP/Z while information flux almost subsumes surprisal on MV/RR constructions. On forward reading time, information flux subsumes surprisal on MV/RR and surprisal subsumes $\max_t D'(t)$ on NP/S.

B Study 2: Additional analysis

Table 4 shows our partial r and p values for a Natural Stories joint linear mixed effect model with

Model	Predictor	r
<i>Forward reading time</i>		
MV/RR	Surprisal	.004
	$\max_t D'(t)$	.147*
NP/S	Surprisal	-.099
	$\max_t D'(t)$	.068
NP/Z	Surprisal	.255***
	$\max_t D'(t)$	.036
<i>Gaze duration</i>		
MV/RR	Surprisal	.082
	$\max_t D'(t)$	.094
NP/S	Surprisal	.023
	$\max_t D'(t)$	-.055
NP/Z	Surprisal	.251***
	$\max_t D'(t)$	-.047
<i>Go-past time</i>		
MV/RR	Surprisal	-.014
	$\max_t D'(t)$	.014
NP/S	Surprisal	.070
	$\max_t D'(t)$	.142***
NP/Z	Surprisal	.102
	$\max_t D'(t)$	-.089
<i>First-pass regressions out (z)</i>		
MV/RR	Surprisal	0.190
	$\max_t D'(t)$	-1.038
NP/S	Surprisal	0.429
	$\max_t D'(t)$	1.115
NP/Z	Surprisal	1.585
	$\max_t D'(t)$	-1.603
<i>Total reading time</i>		
MV/RR	Surprisal	.025
	$\max_t D'(t)$	.088***
NP/S	Surprisal	.062
	$\max_t D'(t)$	.020
NP/Z	Surprisal	-.017
	$\max_t D'(t)$	-.016

Table 3: Partial r for surprisal and $\max_t D'(t)$ in the joint model (surprisal + $\max_t D'(t)$), fit separately per garden-path construction (MV/RR, NP/S, NP/Z) on both critical and spillover words. For first-pass regressions out (logistic models), z statistics are reported in place of r . Significance: * $p < .1$, ** $p < .05$, *** $p < .01$, **** $p < .001$.

Model	Predictor	r
Surprisal	Surprisal	.717****
$\max_t D'(t)$	$\max_t D'(t)$	.524****
Surprisal + $\max_t D'(t)$	Surprisal	.723****
	$\max_t D'(t)$	-.078

Table 4: Partial r by predictor on Natural Stories. Significance: * $p < .1$, ** $p < .05$, *** $p < .01$, **** $p < .001$.

both predictors to analyze which carries more information when sharing variance. In contrast to our results on Study 1, the subsumption appears reversed. p -values for surprisal remain statistically significant regardless of whether it is in a solo or joint model, however $\max_t D'(t)$ becomes insignif-

icant when in the combined model. $\max_t D'(t)$'s r value also drops while surprisal's rises slightly. Surprisal also maintains the direction of correlation while $\max_t D'(t)$ flips.

C Collinearity and Shared Variance

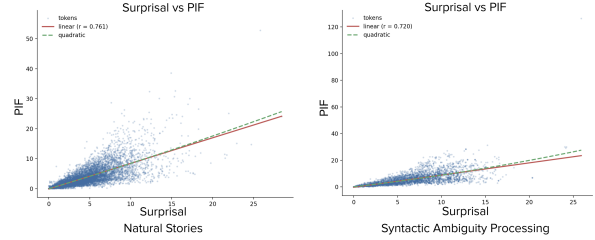


Figure 5: Regression plots between surprisal and peak information flux on Syntactic Ambiguity Processing and Natural Stories.

Surprisal and peak information flux share substantial variance, with a Pearson's r of 0.7203 on SAP, corresponding to a VIF of 2.08, and a Spearman's ρ of 0.8627. Most of the difference appears to be concentrated in higher surprisal bands, with the standard deviation of residuals increasing by surprisal quartile (Q1 = 1.05, Q2 = 1.24, Q3 = 2.74, Q4 = 5.63). These results remain similar for Natural Stories, our second study ($r = 0.7602$, $\rho = 0.8744$; Q1 = 0.37, Q2 = 0.83, Q3 = 1.66, Q4 = 3.93). These quartile progressions show five- to ten-fold increases, depending on dataset, suggesting that the majority of shared variance lives in lower surprisal values. This fan effect can be seen in Figure 5. Our hypotheses regarding peak information flux having sensitivity to reanalysis and reinterpretation are consistent with the standard deviations of residuals at different quartiles of surprisal, where surprisal and peak information flux show the greatest dissociation at higher quartiles.

D Uniform Information Density Comparison

We formalize uniform information density as the squared deviation from the mean surprisal of the sentence, or $UID(w) = (s(w) - \mu_{s(W)})^2$, where w is the word, W is all words in word w 's sentence, and $s(w)$ is the GPT-2 surprisal. We exclude the position 1 of the sentence or story from the calculation of $\mu_{s(W)}$. This was inspired by Meister et al. (2021). We exhibit a correlation coefficient of 0.239 between peak information flux and UID on SAP and 0.396 on Natural Stories.

Construction	N	BIC		ΔBIC
		UID	$\max_t D'(t)$	
<i>Forward reading time</i>				
MV/RR	985	12657.58	12653.84	3.74
NP/S	1,353	17076.78	17079.33	-2.55
NP/Z	1,060	13453.32	13457.71	-4.39
<i>Gaze duration</i>				
MV/RR	1,874	24076.75	24073.60	3.14
NP/S	1,894	23981.11	23980.83	0.27
NP/Z	1,864	23703.72	23709.50	-5.78
<i>Go-past time</i>				
MV/RR	1,865	30585.64	30585.77	-0.13
NP/S	1,886	28637.35	28635.18	2.16
NP/Z	1,858	30387.17	30387.29	-0.13
<i>First-pass regressions out</i>				
MV/RR	1,864	2588.49	2587.54	0.95
NP/S	1,897	2266.62	2265.83	0.79
NP/Z	1,860	2554.70	2556.50	-1.80
<i>Total reading time</i>				
MV/RR	2,014	31640.76	31629.19	11.57
NP/S	2,067	30606.07	30610.14	-4.06
NP/Z	2,046	31388.48	31389.47	-0.99

Table 5: BIC for UID and $\max_t D'(t)$ models on SAP with the same calculation pipeline as the table from Study 1 but running uniform information density instead of surprisal.

Predictor	N	BIC
UID	800,480	9481520.85
$\max_t D'(t)$	800,480	9479998.00
ΔBIC		1522.85

Table 6: BIC for UID and $\max_t D'(t)$ models on Natural Stories with the same calculation pipeline as the table from Study 2 but running uniform information density instead of surprisal.

E Inter-construction Analyses

These figures show the inter-construction results for magnitude asymmetry and time of information flux. They're consistent with the general garden path effect pattern of MV/RR and NP/Z having greater garden path effects than NP/S. Magnitude asymmetry increases the most across quintiles in MV/RR and NP/Z constructions while increasing more modestly in NP/S, and time of information flux is greater in MV/RR and NP/Z relative to NP/S. This is consistent in both critical and spillover regions.

F Hyperparameters

We use a $\lambda_{\max}$ and timestep count t of 10 and 100 for Syntactic Ambiguity Processing and 30 and 300 for Natural Stories. These were chosen based

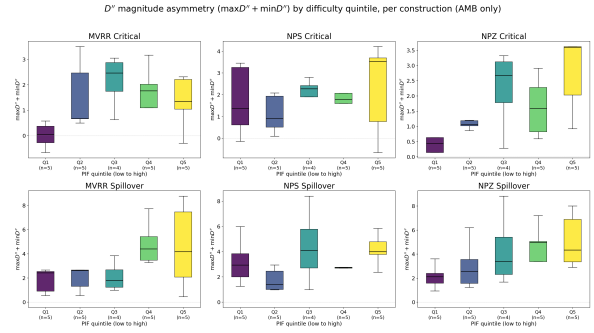


Figure 6: Magnitude asymmetry plots split by construction.

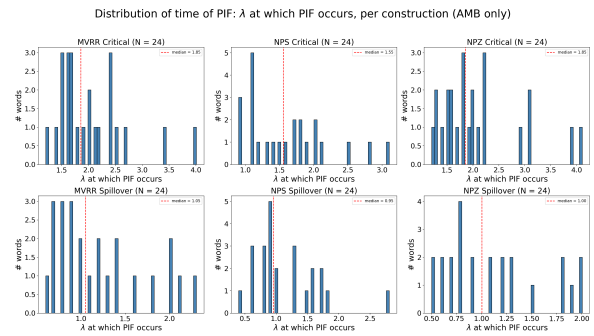


Figure 7: Time of information flux histograms split by construction.

on the total amount of required steps to achieve timecourse convergence, however our results don't show much dependence on these presuming they're at or above a reasonable threshold, such as 10 and 100. We choose a candidates count of $n = 100$ per candidate type in order to provide good coverage of each type, and this means the maximum amount of candidates per word can be 301 since candidate sets are required to include the target word. We also remove external punctuation from targets when calculating peak information flux so as to score lexical base, however our results don't seem to show much dependence on this either. This has no numerical influence on the SAP study specifically since none of the ambiguous garden path stimuli contain punctuation at the critical or spillover regions. All punctuation is maintained in the model's context.

G Hardware

We used a single NVIDIA RTX 4090 GPU with 24GB of VRAM and an Intel i9-13900KF CPU with 64GB of RAM to generate our results. The number of GPU hours used is estimated to be less than 200.