Readers make targeted regressions to plausible errors in reanalysis of "noisy-channel garden-path" sentences

Thomas Hikaru Clark **Roger Levy** **Edward Gibson**
MIT Brain and Cognitive Sciences
43 Vassar Street, Cambridge MA, USA
Correspondence: thclark@mit.edu

Abstract

A key question in psycholinguistics is how inferences about the meaning of linguistic input unfold incrementally a comprehender's mind. In this work, we study reading dynamics for "noisy-channel garden-path" sentences, which temporarily appear well-formed but feature late-appearing violations of expectation that can be resolved not by inferring an alternative syntactic structure, but by inferring the presence of an error. We find evidence for targeted regressions – eye movements towards regions that are promising loci of possible errors in light of later-arriving information, showing patterns consistent with the posterior inferences of a model of noisy-channel processing with reanalysis. We discuss the implications of these findings for theories of noisy-channel language comprehension and information-theoretic explanations of reading dynamics.

1 Introduction

A key question in psycholinguistics is how inferences about the meaning of linguistic input unfold over time in the mind of a comprehender. Natural language contains **ambiguity**: a given input can be compatible with multiple latent grammatical structures (e.g., *I watched the hiker with the binoculars*), sound sequences can be ambiguous between different lexical items, and pronouns can be ambiguous between different referents. Language comprehension is also constrained by **incrementality**: linguistic input is typically processed unit by unit in a linear sequence. Under strict incrementality, information about later words in a sentence cannot influence the initial processing of earlier words. To handle ambiguity, an incremental comprehender needs to either maintain uncertainty regarding an ambiguous word or structure until disambiguating information is reached, or commit to one interpretation, even if it ultimately proves incorrect.

Sentences in which incremental processing is led astray tend to elicit processing difficulties in humans; these examples provide insights into the nature of sentence comprehension. In **garden-path sentences**, an initially likely incremental syntactic parse of a sentence is rendered incompatible by later disambiguating material, potentially forcing comprehenders to **reanalyze** the sentence as having an alternative structure. An open question is whether this garden-path effect extends not just to syntactic reanalysis, but to a broader notion of error-aware reanalysis, where readers take into account possible errors and noise in the linguistic signal, in keeping with **noisy-channel** theories of comprehension (Levy, 2008b; Gibson et al., 2013).

To probe the nature of reanalysis during comprehension, **regressive eye movements** are particularly informative. During reading, readers do not simply advance their gaze sequentially from word to word, but often revisit earlier words, suggesting a shift in attention to, and possible reanalysis of, previously processed material. In this study, we investigate whether readers make targeted regressions to the locations of explainable errors. Distinct from syntactic analysis, this evaluates whether reading behavior is sensitive to possible production errors in the linguistic input, a prediction of noisy-channel theories. This paper attempts to answer two research questions. **Q1**: Does reanalysis during language processing extend to possible errors and not just syntactic ambiguity? **Q2**: Can an algorithmic-level model of noisy-channel language processing predict patterns of reading regressions for anomalous sentences? To foreshadow our findings, we find evidence for targeted regressions to the location of likely word substitution errors, and demonstrate that this is consistent with the posterior predictions of a noisy-channel inference model.

1.1 Incremental processing and garden-paths

Much of the research on reading behavior in psycholinguistics has been focused on incremental processing. According to surprisal theory (Hale,

2001; Smith and Levy, 2013; Levy, 2008a), the processing difficulty of a given linguistic unit, e.g. a word, is proportional to its surprisal (or negative log probability) in context. Many experimental reading paradigms, such as self-paced reading and the Maze task, only provide incremental reading measures (Aaronson and Scarborough, 1976; Mitchell and Green, 1978; Forster et al., 2009).

Classic garden-path sentences provide a clear example of incremental processing difficulty induced by the ambiguity present in language (Bever, 1970; Frazier, 1979; Paape and Vasishth, 2022b). For example, in the sentence *The bird perched on the branch sang sweetly*, the word *perched* initially invites being parsed as the main verb; however, the word *sang*, when reached, rules out this parse. Crucially, the existence of a garden-path effect suggests that comprehenders do not necessarily track all possible grammatical structures consistent with a given input prefix, but may instead greedily commit to an initially likely hypothesis, resulting in detectable slowdowns when disambiguating material is encountered.

1.2 Regressive eye movements in reading

Human reading behavior is not always strictly incremental. Regressive eye movements, or **regressions**, are common — it is estimated that 5-20% of saccades are regressive (Rayner, 1998). While some regressions may be driven by low-level spatial constraints, such as the physical layout of text (Mitchell et al., 2008), preventing readers from making regressions was shown to harm comprehension (Schotter et al., 2014); meanwhile, garden-path sentences elicit more regressions than control materials, which supports the idea that structural ambiguity leads to re-reading.

The Selective Reanalysis hypothesis (Frazier and Rayner, 1982) proposes that reading regressions represent readers' attempts to reanalyze the syntactic structure of earlier material. Under this account, readers can make errors in incremental parsing by committing to the wrong syntactic parse, which causes later violations of expectations; readers then tend to regress to the location of the parsing error to re-parse the input. This provides one explanation of why reading regressions are especially likely to initiate at the disambiguating point of garden-path sentences. Recent work, however, has questioned the validity of the Selective Reanalysis Hypothesis, finding that garden-path sentences were often misunderstood and that re-reading did not improve comprehension (Christianson et al., 2024). Paape et al. (2021) found only weak and inconclusive support for targeted regressions to critical context words in sentences with late-breaking anomalies, and Paape and Vasishth (2022a) found support for confirmatory but not reanalytical regressions in syntactic garden-path sentences, using a bidirectional self-paced reading paradigm. Meanwhile, a recent large-scale eye-tracking study of re-reading behavior by Timkey et al. (2025) found evidence for targeted regressions to critical regions in syntactically challenging sentences.

Wilcox et al. (2024b) perform an information-theoretic analysis of regressions, and find that high **pointwise mutual information** (PMI) between pairs of words in a sentence are associated with regressions during reading. High PMI signifies that two words are predictive of each other (e.g. *I gave the **dog** a **bone***), while low PMI signifies that two words tend to appear together *less* than one would expect by chance (e.g. *I gave the **dog** a **knife***). This result is taken as evidence for a **reactivation account** of regressions, as opposed to a **reanalysis account**: readers tend to revisit words which are information-theoretically congruent, not incongruent, with the current word. The reactivation account is consistent with work finding that reading regressions are predicted by syntactic dependencies within a sentence (Lopopolo et al., 2019), which tend to exist between pairs of words with high mutual information (Futrell, 2019).

A question left open by past work is how regressions manifest in sentences that contain production errors (e.g. word substitutions), as opposed to syntactic ambiguity. Given that errors in printed texts are relatively rare, one might expect to find little evidence for the reanalysis account in reading of naturalistic language (the focus of most past studies), or that this evidence would be dominated by the much more common case of regressions for reactivation — a possibility directly raised directly by Wilcox et al. (2024b). Word substitutions, for example, preserve dependency structure while creating pairs of words in a dependency relation which nevertheless have low PMI; it is currently unclear whether human readers would be more or less likely to make regressive eye movements between such a pair of words.

1.3 Algorithmic accounts of noisy-channel language processing

The noisy-channel theory of language processing proposes that comprehenders arrive at non-veridical interpretations of sentences when more plausible alternatives exist, rationally integrating both the prior probability of different intended sentences and the error likelihood (Levy, 2008b; Gibson et al., 2013). Recent work has modeled how sentence processing for anomalous linguistic input may unfold over time for sentences with strong violations of incremental expectations, like *The storyteller could turn any story into an amusing* ***antidote*** (Ryskin et al., 2021; Li and Ettinger, 2023; Li and Futrell, 2024).

Clark et al. (2025a) propose a model with specific algorithmic commitments regarding both the incremental processing and reanalysis of possibly noisy utterances. In this model, Sequential Monte Carlo (SMC) inference is employed to approximate the posterior distribution over intended messages, conditional on an observed noisy string. SMC is an approximate Bayesian inference algorithm for sequential observations, which naturally instantiates a tradeoff between the number of particles and the exactness of inference: as the number of particles approaches infinity, the SMC approximation approaches the true posterior (Lew et al., 2023; Doucet et al., 2001; Naesseth et al., 2024). In past work, the number of SMC particles is used as a proxy for cognitive resources (Levy et al., 2008; Clark et al., 2025b). This computational approach therefore provides an explanation for how human comprehenders can be approximately rational, despite the intractability of computing exact posterior inferences regarding a large space of alternative interpretations of a noisy sentence.

This model is by default incremental, processing words in sequential order and updating the posterior distribution over latent variables at each time step. At the same time, the model supports reanalysis of earlier commitments via **rejuvenation** moves that propose changes to earlier commitments (e.g., inferring that an earlier word was actually an error). Unlike Levy et al. (2008), which treats the input string as veridical and performs inference over the uncertainty in latent syntactic structures, this noisy-channel model treats the input string as possibly non-veridical and performs inference over possible alternative *intended* sentences. In this work, we use this computational model to generate posterior inferences for possibly noisy experimental stimuli, yielding predictions of how the experimental conditions may systematically differ from each other (Section 2.3). In particular, the model identifies words which are likely to be errors in light of later-arriving information, even if they initially appear perfectly well-formed using only previous context.

1.4 "Noisy-channel garden-path" sentences as a key testbed for theories of processing

A useful case for testing theories of reanalysis in reading is a sentence which elicits a garden-path effect that can be resolved not by syntactic reanalysis but by hypothesizing a production error at an earlier part of the sentence. An example of such a **"noisy-channel garden-path"** sentence is *The boy licked the big round ball into the net* (full example in Table 1). When such a sentence is being incrementally processed, the `Predicate` is semantically, rather than syntactically, incongruous with earlier context. This incongruity, paired with the form-based similarity of the word *licked* to the more globally congruous word *kicked*, potentially invites an error-correction inference on the part of a comprehender. Each implausible sentence can be directly compared to a plausible counterpart which is identical up until the `Predicate`.

Critically, no purely incremental processing algorithm (e.g., autoregressive language model surprisal) can make different predictions regarding regressive reading specifically to `CriticalWord` in the matched `Plausible` and `Neighbor-GP` sentences (see Table 1). Incremental language model surprisal *does* predict differences in behavior at the `Predicate`, and one plausible hypothesis is that when a threshold of surprisal is reached, some form of repair or reanalysis is triggered. However, to know specifically *where* a comprehender should focus their re-reading effort, we need an algorithm that instantiates reanalysis. Unlike in traditional garden-path sentences, these noisy-channel garden-paths have a specific location that can be hypothesized as a production error. Three plausible accounts are described below.

A: Purely incremental account. Incremental features capture all explainable variation in reading times. When encountering a semantic violation at the `Predicate`, the reader either covertly performs reanalysis without an eye-movement correlate, or accepts the semantically incongruous but syntactically valid `Predicate` without reanalysis. In either case, we expect a slowdown at the `Predicate`, but

Table 1: Example stimuli — plausible and garden-path sentences.

Condition	Preamble	CriticalWord	Intervening	Predicate
Plausible	The boy	kicked	the big round	ball into the net.
Plausible	The boy	licked	the big round	lollipop with delight.
Neighbor-GP	The boy	licked	the big round	ball into the net.
Neighbor-GP	The boy	kicked	the big round	lollipop with delight.

this account does not predict regressions to earlier words.

B: Surprisal triggers non-targeted re-reading. When surprising material is reached, readers are more likely to initiate a regression to earlier material, without specifically targeting the location of likely errors. A simple version of this account is that readers simply re-read the sentence from the beginning. Under this account, the CriticalWord in the Neighbor-GP condition in Table 1 would not attract more re-reading than other regions in the sentence.

C: Surprisal triggers targeted re-reading. This account predicts that when surprising material is reached, readers are more likely to initiate a regression targeted to the location of likely errors in the sentence. One theory that predicts targeted re-reading is an algorithmic-level model of noisy-channel comprehension, which proposes multiple possible cues that may identify an earlier word in a sentence as being a fruitful location for reanalysis. Broadly, these cues include the linguistic **prior** (how probable a word is in context) and the error **likelihood** (how likely a particular error is, according to a comprehender's mental model of errors). The challenge for a reader is to identify earlier words which (a) have an alternative interpretation that is a near neighbor in the space of errors and (b) this alternative leads to a more plausible (higher-prior) sentence. Additionally, this requires a linking hypothesis, which is that reanalysis of a word is mediated by additional eye movements towards that word after it has been read, and the ensuing additional reading duration.

2 Methodology

We conduct a reading experiment aimed at answering our key research questions. The experimental setup and analysis plan were preregistered via OSF: https://osf.io/qtnxa.

2.1 Materials

We systematically manipulate sentences to test reading behavior for noisy-channel garden path sentences and a variety of control conditions (Tables 1 and 2). All target materials consist of the following regions: Preamble, CriticalWord, Intervening, and Predicate. We generated 36 items, each of which appears in 5 conditions. Additionally, 36 filler items containing no errors were shown. More examples of experimental materials are shown in Appendix B.

Condition	CriticalWord	Predicate
Typo	kjcked	ball into the net.
Typo	ljcked	lollipop with delight.
Unrelated-GP	read	ball into the net.
Unrelated-GP	read	lollipop with delight.
Late-Error	kicked	breath after the run.
Late-Error	licked	breath after the run.

Table 2: Additional control stimuli for one item. All variants of an item share the same Preamble and Intervening material: "The boy" and "the big round".

Within each condition, items are counterbalanced by having two versions of CriticalWord; this leads to 2 variants of each item per condition. The Neighbor-GP condition is formed from the Plausible condition by simply swapping the Predicate regions of the two variants. This leads to a tightly controlled comparison where lexical items are completely balanced, with plausibility depending only on the identity of the Predicate. For example, both prefixes *The boy licked the big round...* and *The boy kicked the big round...* appear in both the Plausible and Garden-Path conditions, depending only on the value of Predicate (*ball into the net* or *lollipop with delight*). Crucially, CriticalWord in the two variants are orthographic neighbors of each other, e.g. *lick* and *kick*, providing a way of assessing sensitivity to error likelihood.

The Unrelated-GP condition provides a comparison where the CriticalWord region is an unrelated (non-neighbor) lexical item. These sentences can also be interpreted by positing a word substitution error at the CriticalWord region, but unlike in the Neighbor-GP condition, the orthographic form of CriticalWord provides no cue towards a

more plausible word. In the case of Typo errors, CriticalWord is a non-word. This means that the sentence does not contain any temporary ambiguity, and thus no garden-path effect. In the Late-Error condition, Predicate is incoherent with the earlier parts of the sentence. Thus, while the material up to the Predicate is matched by sentences in the Plausible condition, the sentence resists reanalysis via a simple word substitution error.

2.2 Reading time data collection

To gather reading regression time data experimentally, we employ the Mouse Tracking for Reading (MoTR) paradigm (Wilcox et al., 2024a). In other reading paradigms, such as unidirectional self-paced reading (SPR) and Maze, regressions are not possible (Aaronson and Scarborough, 1976; Mitchell and Green, 1978; Futrell et al., 2021; Forster et al., 2009). Meanwhile, bi-directional SPR found evidence for selective re-reading only in very difficult syntactic garden-path sentences (Paape and Vasishth, 2022c,a). In MoTR, participants move their mouse to un-blur a small "spotlight" region within a blurry text. By tracking the position of the mouse on the screen, experimenters can compute how long the mouse lingers over each word as a proxy for reading time. After each sentence is read, participants select a response from one of the following choices: "Sentence was OK", "I noticed an error", or "Not sure". A breakdown of responses by condition is shown in Appendix A.

Participants We recruited 200 participants via the online platform Prolific, in accordance with an existing IRB protocol at the authors' institution. Participants were self-reported native English speakers residing in the United States, and were paid $4. The task took approximately 15 minutes. Participants provided informed consent before beginning the experiment. Participants were pseudo-randomly assigned to one of 10 experimental lists, which rotated each item through the 10 variants; each participant thus saw each of the 36 items in exactly 1 condition, and each of the 360 unique item variants was seen by 20 participants. Experimental items were interspersed with 36 filler sentences containing a variety of syntactic constructions and semantic topics, which were seen by all participants.

Reading measures The MoTR paradigm results in raw data in the form of many samples of mouse positions on the screen during each trial. Using the post-processing pipeline of Wilcox et al. (2024a), we generate the following continuous reading measures: first fixation duration, gaze duration, go past time, right-bounded reading time, and total duration. We also generate the following binary variables: first pass fixation, first pass regression out, and regression in. Regressions into the Predicate (the final region) are still possible, since the reader may move their mouse beyond the end of the sentence and then back into the sentence.

Exclusion criteria We exclude data points based on the following criteria (Wilcox et al., 2024a): We exclude data from individual trials if the participant fixated on fewer than 20% of total words in the sentence. We exclude any words whose gaze duration time was more than 3 standard deviations greater than the mean gaze duration for that word, indicating abnormally long gaze time for that word. We exclude all of a participant's data if they report "I noticed an error" for more than 20% of fillers, which suggests inattentive reading.

2.3 Computational Modeling

We employ the model of Clark et al. (2025a) to produce noisy-channel inferences for each sentence in the study [1]. In this model, $\mathbf{u}$ represents a sequence of observed words, while a set of K weighted particles $\{x_t^{(i)}\}, i = 1 \dots K$, each with weight $w_t^{(i)}$, represents hypotheses about the model state, e.g. the inferred intended sentence and the sequences of errors that map between the intended sentence and the observed sentence. In this model, the prior over strings is derived from the GPT-2 language model, which has been shown to be a strong predictor of human reading times, more so than larger models (Shain et al., 2024; Oh and Schuler, 2023); the model additionally incorporates inferences about possible errors and intended alternatives using Bayesian inference.

Two model-based quantities are of particular interest. First, we extract incremental surprisal, i.e. the negative log probability of an observed word in context, which is approximated by taking the average particle weight at position t:

$$P(\mathbf{u}_t \mid \mathbf{u}_{1:t-1}) = \int P(\mathbf{u}_t \mid x_t) P(x_t \mid \mathbf{u}_{1:t-1}) dx_t \approx \frac{1}{K} \sum_{i=1}^{K} w_t^{(i)}$$

[1] https://github.com/thomashikaru/noisy_channel_model; further details are in Appendix C.

Figure 1 demonstrates the average per-word surprisal for different regions in the experimental materials, separated by condition. The most tightly controlled pair of conditions are Plausible and Neighbor-GP, which have identical surprisal values until the Predicate is reached, at which point Neighbor-GP elicits markedly higher surprisal. The Typo condition is unique in having an immediately obvious error, leading to a large spike in surprisal at the CriticalWord. While this predictor may explain variance in incremental reading behavior or the probability of a regression being initiated at a particular word, it cannot on its own predict targeted regressions to likely errors earlier in the sentence, in light of new information.

Second, the model provides an estimate of the posterior distribution over actions at each word in the input utterance, where the NORMAL action denotes no error, in contrast to the FORM-BASED SUBSTITUTION, MORPHOLOGICAL SUBSTITUTION and SEMANTIC SUBSTITUTION actions (see Figure 4 for inferences for an example sentence). The posterior probabilities for actions at each word provide an estimate of the probability that, in light of the full utterance, a particular word contained a particular type of error, or no error. Figure 1 illustrates the average per-word posterior error probability for different regions, separated by condition. Echoing the pattern of surprisal values, the Typo condition has the largest error probability at the CriticalWord. Crucially, however, despite the identical surprisal values at the CriticalWord region in Plausible and Neighbor-GP, the latter has a higher posterior probability of error at the CriticalWord (and is also higher than in Unrelated-GP and Late-Error). Additional examples of model-generated noisy-channel inferences are provided in Appendix C.

2.4 Hypotheses

We hypothesize that the results will align with the **targeted regressions account**: readers will make targeted regressions to the location of likely errors which are promising loci for reanalysis. Under this hypothesis, the probability of regression out from the Predicate region will be higher in all non-Plausible conditions than in the Plausible condition. Furthermore, the CriticalWord in the Neighbor-GP condition will be associated with a higher probability of regressions in, as well as re-reading time, than in any other condition. We predict that this interaction effect will be greater for Neighbor-GP than for Unrelated-GP, implying that targeted regressions and re-reading preferentially target errors that are more likely, compared to unrelated word substitutions.

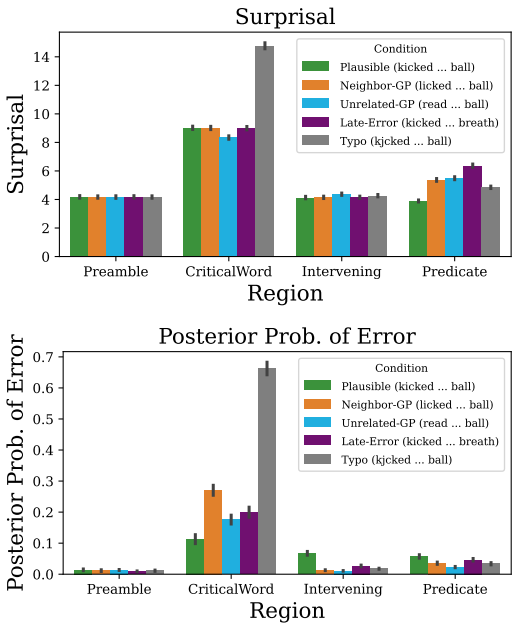


Figure 1: Incremental surprisal and error probability, computed using noisy-channel inference model. Error bars denote 95% CIs.

2.5 Statistical Analysis

Re-reading time: We use a Bayesian mixed-effects hurdle-lognormal model for modeling the continuous re-reading (total minus gaze) duration values. Word frequency, word length, incremental surprisal, word position, and part of speech are used to predict the binary value of whether a word will have zero reading time. These same predictors, and the additional predictor of a Region × Condition interaction, are used to predict re-reading duration when it is non-zero. Treatment coding is used for categorical predictors; Plausible is used as the reference level for Condition, while Intervening is used as the reference level for Region. In each case, the reference level provides a baseline to which the focal manipulations can be compared. Per-participant and per-item random intercepts were included, and per-participant random slopes were included for frequency, length, surprisal, and word position. We note that our preregistration originally did not include re-reading duration as a reading measure, but this was added to capture the specific feature of time spent reading after the first pass.

Regressions in/out: We use a Bayesian mixed-effects logistic regression model, with the same predictors and reference levels as above, to predict regressive eye movements, both for a given source and for a given target (**regressions out** and **regressions into**). Random intercepts and slopes follow the same scheme as for the above models.

3 Results

3.1 Implausible predicates trigger regressions

Inspecting the fixed effects of the model predicting regressions out, we observe an interaction between Condition and Region for `Neighbor-GP` × `Predicate` ($\beta = 0.333$, CrI = [0.196, 0.469]), `Unrelated-GP` × `Predicate` ($\beta = 0.230$, CrI = [0.091, 0.368]), and `Late-Error` × `Predicate` ($\beta = 0.166$, CrI = [0.028, 0.305]) — these conditions all had higher probabilities of regressions out at the `Predicate` than the `Plausible` (baseline) condition. Meanwhile, the `Typo` × `Predicate` interaction shows a lower probability of regressions out at the `Predicate`, compared to the `Plausible` condition ($\beta = -0.450$, CrI = [-0.597, -0.300]), likely because readers tended to already have detected an error earlier in the sentence. Full statistical results are presented in Appendix D. These results align with the patterns observed in Figure 2.

3.2 Likely errors attract regressions in

Inspecting the fixed effects of the model predicting the rate of regressions in, we again observe an interaction between Condition and Region for `Neighbor-GP` × `CriticalWord` ($\beta = 0.369$, CrI = [0.242, 0.499]) and `Unrelated-GP` × `CriticalWord` ($\beta = 0.279$, CrI = [0.156, 0.406]). There was no evidence for an interaction for `Late-Error` × `CriticalWord` ($\beta = 0.005$, CrI = [-0.118, 0.127]) or `Typo` × `CriticalWord` ($\beta = 0.0$, CrI = [-0.202, 0.189]). Notably, the largest interaction was observed for the `Neighbor-GP` condition, where an orthographic neighbor of the `CriticalWord` makes for a much more plausible sentence. We see a positive, but numerically smaller interaction in the `Unrelated-GP` condition, where the `CriticalWord` is still an "odd word out", but there is no obvious mapping from the `CriticalWord` to a more plausible alternative.

Interactions between Condition and Region were also observed in re-reading time (total minus gaze duration), for `Neighbor-GP` × `CriticalWord` ($\beta = 0.215$, CrI = [0.125, 0.310]), `Typo` × `CriticalWord` ($\beta = 0.456$, CrI = [0.333, 0.578]), `Unrelated-GP` × `CriticalWord` ($\beta = 0.101$, CrI = [0.007, 0.196]), and `Late-Error` × `CriticalWord` ($\beta = 0.133$, CrI = [0.037, 0.232]); there was no evidence for other Condition × Region interactions. Full statistical results are presented in Appendix D.

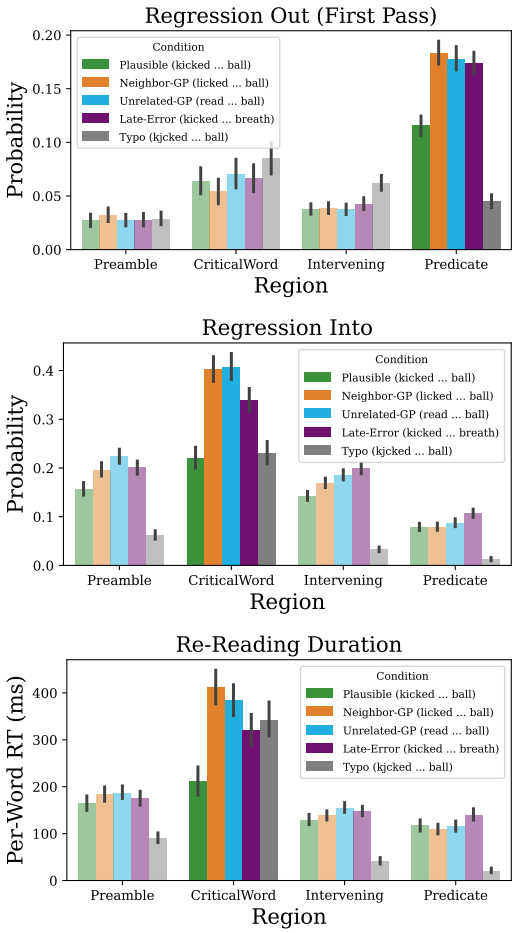


Figure 2: Reading measures as a function of region and condition. Error bars denote 95% CIs.

4 Discussion

4.1 Targeted regressions towards likely errors

Our results suggest that **regressions out** are triggered by violations of expectation, found in surprising predicates. This is consistent with existing models of incremental processing (Levy, 2008a; Hale, 2001; Smith and Levy, 2013). However, what has remained unclear is how readers allocate attention and effort during reanalysis of problematic sentences.

Turning to the target of regressions, our results indicate that readers are more likely to regress into a word in a sentence which becomes a probable

locus of error in light of new information. Despite the Plausible and Neighbor-GP conditions having identical incremental properties up until the Predicate, there are more regressions into and longer re-reading durations on the CriticalWord in Neighbor-GP. Comparing the CriticalWord to other regions in the Neighbor-GP condition, we see that the probability of regressions in and re-reading duration are both higher there than at any other region. This rules out the idea that surprising predicates lead to re-reading of the entire sentence (a reasonable alternative explanation), which would result in uniformly higher duration across the whole sentence and more regressions into the Preamble than any other region. Comparing the different conditions, we observe that Neighbor-GP and Unrelated-GP have a higher probability of regressions into and higher re-reading duration than the other error conditions, with Neighbor-GP having a larger interaction estimate than Unrelated-GP. This suggests that reanalysis is sensitive to locations in a sentence where correcting a single error can repair the sentence into something more plausible, especially when this error is a likely one (e.g. an orthographic neighbor).

This targeted behavior could be explained by readers needing time to assess the probability of a language producer making a given error, or because they have uncertainty about the fidelity of their own perceptual input and want to verify or "double-check" what the sentence actually said. For obvious typos, on the other hand, participants' reading behavior is markedly different — they tend to linger on it on the first pass, but are less likely to revisit the error. The fact that readers show special treatment of typos and orthographic neighbors suggests that readers may employ a mental model of possible errors when interpreting uncertain input (Gibson et al., 2013). Our findings are complementary to work finding evidence for targeted regressions that correspond to syntactic reanalysis (Timkey et al., 2025), showing that an analogous process may be at play in reanalyzing earlier commitments during comprehension as *errors* in light of late-arriving information.

4.2 Algorithmic accounts of reanalysis

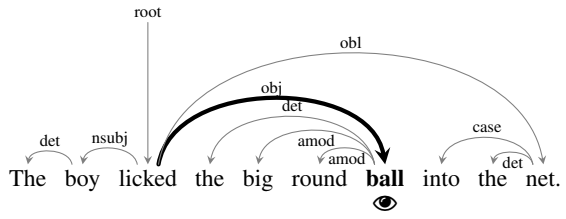
Existing accounts of eye movements during reading have successfully been applied to different levels of granularity, from the timescale of individual saccades (Legge et al., 1997; Reichle et al., 2003; Bicknell et al., 2020), to high-level information-seeking behavior (Gruteke Klein et al., 2024). Yet edge-case behavior, like the reanalysis of earlier words as errors, has not been explicitly modeled in these works. Meanwhile, works like Levy et al. (2008) model processing of syntactic ambiguity and garden-path sentences with resource-rational approximate inference algorithms, but without a direct linking hypothesis to reanalytical reading behavior when inference fails. Clark et al. (2025a) apply a similar class of inference algorithms to the processing of "noisy" sentences, where rejuvenation proposals simulate reanalysis of earlier words as errors. This model, however, is under-specified with regard to the "control flow" governing the sequential order of inference steps during reading, and supports both unconditional re-reading of a sentence or conditionally triggered reanalysis. In the model, conditional reanalysis instantiates the notion that readers decide whether to reanalyze based on an estimate of the probability that the sentence contains an error, estimated using the surprisal of the current word, normalized by its unigram probability (i.e. "How surprising is this word *in* context, relative to how surprising it is *out of* context?"). This is closely related to metrics such as SLOR (Kann et al., 2018) and MORCELLA (Tjuatja et al., 2025). Future work can more precisely establish how algorithmic details within approximate probabilistic inference may predict human eye movements, including individual variation.

In this study, predicates in the Neighbor-GP condition were *not* associated with additional reading duration, compared to minimally different Plausible sentences, but rather with a higher probability of regressions out. This suggests that when re-reading is an option, it is triggered quite readily. In contrast, paradigms which do not allow for re-reading (e.g., self-paced reading) have been shown to instead induce slowdowns in spillover regions; notably, this may not be the optimal reading strategy for readers from the perspective of successful comprehension (Schotter et al., 2014). A takeaway for theories of human language processing is that purely incremental processors are insufficient to capture human behavior, and non-optimal from the perspective of handling the uncertainty, ambiguity, and noise present in naturalistic language usage. Within the approximate probabilistic inference framework for modeling noisy-channel comprehension, reanalysis is strongly motivated as a resource-rational, efficient strategy: to guarantee recovery from errors in purely incremental processing would

require a massive amount of parallel computation and memory to encode all possible alternatives of a linguistic unit, just in case this alternative is later rendered likely by new information. In contrast, allowing the model to execute rejuvenation moves only when an "error signal" is detected would allow processing to proceed with far fewer cognitive resources. Since errors are generally infrequent compared to non-errors, but are crucial to correct when they occur, a flexible strategy capable of deploying reanalysis when needed is an efficient solution to the possibly-noisy language comprehension problem. More broadly, this framework has implications for the processing of syntactic garden-path sentences as well: many garden-path sentences are amenable to alternative interpretations in terms of missing or substituted words rather than syntax (e.g. positions A and B in *The bird [that]$_A$ perched on the branch [and]$_B$ sang sweetly*).

4.3 Syntactic dependencies and PMI

Past work has suggested that reading regressions align with both the latent syntactic dependency structure of a sentence (Lopopolo et al., 2019) as well as with the PMI between pairs of words in a sentence (Wilcox et al., 2024b); dependency relations are also known to be associated with high PMI (Futrell, 2019). Our materials, which manipulate the plausibility of a sentence within each item while preserving the same syntactic structure, provide a way to disentangle the influence of dependency structure and PMI. Appendix E reports a systematic comparison of the PMI between `CriticalWord` and `Predicate` across conditions; as would be expected, `Plausible` items have significantly higher PMI between these two regions than `Neighbor-GP` items. The dependency structure of an example garden-path sentence is shown below, with a bolded arrow indicating a dependency arc between `Predicate` and `CriticalWord`:



Our results show that previously read words that are in a given dependency relation are *more* likely to be regressed to when source and target are low-PMI (*licked, ball*) than when they are high-PMI (*licked, lollipop*). This provides evidence for re-analytical regressions (as opposed to reactivation regressions), which was inconclusive in the naturalistic reading time corpus study of Wilcox et al. (2024b). In naturalistic corpora, high PMI might predict regressions better than low or negative PMI in aggregate, while direct experimental manipulation shows that violations of expectation can drive regressions along dependency arcs when raw PMI would not predict this.

One interpretation of this pattern is that readers maintain a hierarchical mental representation corresponding to dependencies between words, independently of the statistical co-occurrence patterns of words; thus a reader knows that in the example above, the word *licked* is the head of *ball*, and makes a regression from dependent to head to attempt to resolve the violation of expectation. Another interpretation is that the reader may have access to other predictions about the upcoming word, e.g. in the sentence above, a reader may may predict *lollipop* to follow *licked the big round...*; upon seeing the unexpected word *ball*, the reader regresses to the word in the sentence that has high PMI with the predicted but unseen word. The latter account preserves the usefulness of PMI as a proxy for dependency relations, but requires considering the PMI between *predicted* words and earlier words, rather than only that of observed words.

5 Conclusion

In this study, we collected a large dataset comprising mouse-tracking data from 200 participants on 360 unique experimental sentences. Using these data, we present empirical results showing re-reading behavior that is targeted to the locus of plausible errors. This behavior is qualitatively consistent with the condition-level differences in posterior error probability of an algorithmic model of noisy-channel processing, which performs explicit reanalysis to update beliefs about earlier commitments in light of new information. Building on past work on information-theoretic and rational-inference approaches to language comprehension, we provide an integrated explanation of regressive readings during reanalysis, broadly construed — not just syntactic reanalysis, but reanalysis of earlier materials as possible errors. These results demonstrate the flexibility of readers in forming interpretations of noisy language input, and the fine-grained processes underlying robust comprehension during reading.

Limitations

One limitation of this study is the divergence from a truly naturalistic reading environment. Readers may be adopting strategies to answer the questions with minimal effort (for example, skipping the rest of the sentence as soon as a typo is read). Although this is a limitation, it also is a source of insights: in comparison to the Typo conditions, we can be relatively confident that participants are actually reading through the other conditions attentively, and our data show a glimpse into the timecourse of inferences relating to whether a sentence contains an error or not. In the case of non-word errors, this is an easy task, but in sentences with semantic or lexical anomalies, readers make reading regressions to interrogate earlier parts of a sentence before making a final decision.

Individual variation between readers, who may adopt qualitatively different strategies for handling anomalous inputs, is likewise not fully investigated in this study. Data like those collected in this study, which show fine-grained reading behavior for a range of plausible and implausible sentences, provide the potential for new insights towards this question. The pattern of regressions, including the conditions which trigger a regression from a source to a target word, can inform the inference control flow in resource-rational models.

Additionally, this study's exclusive focus on English limits the generality of its findings, and future work can consider typologically diverse languages to gain insight into reanalysis strategies for reading in languages with significantly different writing systems, word order, or morphological complexity.

References

Doris Aaronson and Hollis S. Scarborough. 1976. Performance theories for sentence coding: Some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance*, 2(1):56–70.

Thomas Bever. 1970. The Cognitive Basis for Linguistic Structures. pages 279–352.

Klinton Bicknell, Roger Levy, and Keith Rayner. 2020. Ongoing Cognitive Processing Influences Precise Eye-Movement Targets in Reading. *Psychological Science*, 31(4):351–362.

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.

Kiel Christianson, Jack Dempsey, Anna Tsiola, Sarah-Elizabeth M. Deshaies, and Nayoung Kim. 2024. Retracing the garden-path: Nonselective rereading and no reanalysis. *Journal of Memory and Language*, 137:104515.

Thomas Clark, Jacob Hoover Vigly, Edward Gibson, and Roger Levy. 2025a. A Model of Approximate and Incremental Noisy-Channel Language Processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47(0).

Thomas Hikaru Clark, Jacob Hoover Vigly, Edward Gibson, and Roger P. Levy. 2025b. 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, Suzhou, China. Association for Computational Linguistics.

Arnaud Doucet, Nando de Freitas, and Neil Gordon. 2001. An Introduction to Sequential Monte Carlo Methods. In Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors, *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, pages 3–14. Springer, New York, NY.

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.

Lyn Frazier. 1979. On Comprehending Sentences: Syntactic Parsing Strategies. *ETD Collection for University of Connecticut*.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France. Association for Computational Linguistics.

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(1):63–77.

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.

Keren Gruteke Klein, Yoav Meiri, Omer Shubi, and Yevgeni Berzak. 2024. The Effect of Surprisal on Reading Times in Information Seeking and Repeated Reading. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 219–230, Miami, FL, USA. 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*.

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.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-Level Fluency Evaluation: References Help, But Can Be Spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

Gordon E. Legge, Timothy S. Klitz, and Bosco S. Tjan. 1997. Mr. Chips: An ideal-observer model of reading. *Psychological Review*, 104(3):524–553.

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

Roger Levy. 2008b. A Noisy-Channel Model of Human Sentence Comprehension under Uncertain Input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Honolulu, Hawaii. Association for Computational Linguistics.

Roger Levy, Florencia Reali, and Thomas Griffiths. 2008. Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Alexander K. Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K. Mansinghka. 2023. Sequential Monte Carlo Steering of Large Language Models using Probabilistic Programs. *Preprint*, arXiv:2306.03081.

Jiaxuan Li and Allyson Ettinger. 2023. Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. *Cognition*, 233:105359.

Jiaxuan Li and Richard Futrell. 2024. An information-theoretic model of shallow and deep language comprehension. *Preprint*, arXiv:2405.08223.

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 [cs.CL].

Alessandro Lopopolo, Stefan L. Frank, Antal van den Bosch, and Roel Willems. 2019. Dependency Parsing with your Eyes: Dependency Structure Predicts Eye Regressions During Reading. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 77–85, Minneapolis, Minnesota. Association for Computational Linguistics.

D. C. Mitchell and D. W. Green. 1978. The Effects of Context and Content on Immediate Processing in Reading. *Quarterly Journal of Experimental Psychology*, 30(4):609–636.

Don C. Mitchell, Xingjia Shen, Matthew J. Green, and Timothy L. Hodgson. 2008. Accounting for regressive eye-movements in models of sentence processing: A reappraisal of the Selective Reanalysis hypothesis. *Journal of Memory and Language*, 59(3):266–293.

Christian A. Naesseth, Fredrik Lindsten, and Thomas B. Schön. 2024. Elements of Sequential Monte Carlo. *Preprint*, arXiv:1903.04797.

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. Place: Cambridge, MA Publisher: MIT Press.

Dario Paape and Shravan Vasishth. 2022a. Conscious rereading is confirmatory: Evidence from bidirectional self-paced reading. *Glossa Psycholinguistics*, 1(1).

Dario Paape and Shravan Vasishth. 2022b. Estimating the True Cost of Garden Pathing: A Computational Model of Latent Cognitive Processes. *Cognitive Science*, 46(8):e13186.

Dario Paape and Shravan Vasishth. 2022c. Is reanalysis selective when regressions are consciously controlled? *Glossa Psycholinguistics*, 1(1).

Dario Paape, Shravan Vasishth, and Ralf Engbert. 2021. Does Local Coherence Lead to Targeted Regressions and Illusions of Grammaticality? *Open Mind*, 5:42–58.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26(4):445–476.

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.

Elizabeth R. Schotter, Randy Tran, and Keith Rayner. 2014. Don't Believe What You Read (Only Once): Comprehension Is Supported by Regressions During Reading. *Psychological Science*, 25(6):1218–1226.

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.

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. 2025. Eye movements reveal a dissociation between prediction and structural processing in language comprehension.

Lindia Tjuatja, Graham Neubig, Tal Linzen, and Sophie Hao. 2025. What Goes Into a LM Acceptability Judgment? Rethinking the Impact of Frequency and Length. *Preprint*, arXiv:2411.02528.

Ethan Gotlieb Wilcox, Cui Ding, Mrinmaya Sachan, and Lena Ann Jäger. 2024a. Mouse Tracking for Reading (MoTR): A new naturalistic incremental processing measurement tool. *Journal of Memory and Language*, 138:104534.

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

A Readers' sentence ratings differentiate categories of possible errors

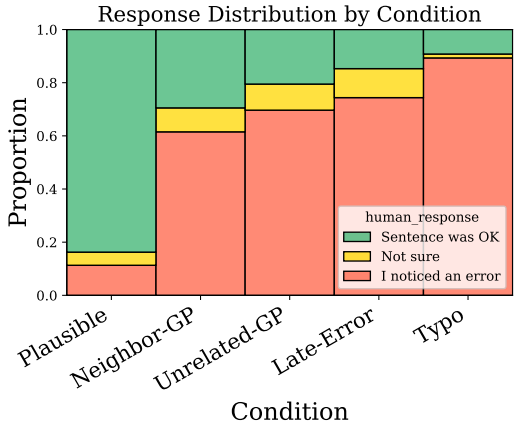


Figure 3: Participant response by condition.

After reading each sentence, participants were asked to select a response regarding whether the sentence contained an error, with the options of "Sentence was OK", "I noticed an error", and "Not sure". Participants were most likely to report an error for the `Typo` condition, followed by `Late-Error`, then `Unrelated-GP`, then `Neighbor-GP`, then Plausible (Figure 3). This provides evidence that participants were performing the task properly. Additionally, we observe that the `Typo` condition led to very little uncertainty in judgments and very low rates of acceptability.

B Additional Experimental Materials

Table 3: Experimental sentences showing `CriticalWord` and `Predicate` manipulations. `Neighbor-GP` sentences result when `Predicate` and `CriticalWord` are mismatched, while Plausible sentences result from matched pairs. Other conditions are not shown.

Sentence
The boy **kicked**/*licked* the big round **ball into the net**/*lollipop with delight*.
His **band**/*hand* was apparently no longer **performing live**/*bandaged tightly*.
In the **base**/*vase* were a wide range of **soldiers from all over**/*flowers from my mom*.
She got a lovely **fan**/*tan* while spending some time in the **gift shop this morning**/*sun over the holidays*.
The **car**/*care* that was provided by the **dealer was excellent**/*doctor was excellent*.
The doctor **wired**/*wiped* the heavily sedated patient's **jaws shut during the surgery**/*blood off during the surgery*.
Can you please **dip**/*zip* that big black **brush into the paint**/*bag up*?
The student **scored**/*scared* a very difficult **goal with a header**/*teacher with a prank*.
The **tires**/*fires* were quickly and efficiently **pumped up by the mechanic**/*extinguished by the firefighters*.
The **bears**/*beers* were last seen **roaming in the forest**/*chilling in the freezer*.
The driver **bumped**/*pumped* the bright red **traffic cone with his car**/*tank full of gas*.
Why did you **shave**/*shove* your really nice **beard for no reason**/*teammate for no reason*?

Sentence
The **checks**/*chicks* were unexpectedly **cashed on Monday**/*hatched on Monday*.
I watched as a **tune**/*tube* was quite masterfully **played by the pianist**/*installed by the plumber*.
The doctor sees patients with **mental**/*metal* or other kinds of **illnesses affecting them**/*fragments in their bodies*.
With two strong **wings**/*wins* this very powerful **bird can fly for hours**/*team can qualify for the finals*.
These **bills**/*hills* definitely appear to be **counterfeit, unfortunately**/*steep, unfortunately*.
The **facts**/*faces* and other details of the **case were published**/*people were blurred*.
The rules state that the **loser**/*lower* of the two relevant **matches would be eliminated**/*amounts would be considered*.
The **garbage**/*garage* definitely needs to be **tossed out due to its smell**/*renovated due to its leaks*.
This **wind**/*wine* is most definitely **blowing from the west**/*delicious with steak*.
Thank you for **taking**/*making* such good **care of her yesterday**/*pancakes for breakfast*.
After being **seated**/*sealed* carefully in the small **theater, the audience fell silent**/*jar, the jam was sold*.
Because of all the **treats**/*threats* my dog has become very **overweight and unhealthy**/*scared and timid*.
The **chart**/*cart* contained lots of **data and statistics**/*fruits and vegetables*.
These **saints**/*stains* had been talked about and **venerated by the worshippers**/*removed by the cleaners*.
She knew how to **untie**/*unite* every single one of the **knots for the sailboat**/*community members*.
Are you able to **reverse**/*reserve* both of these **transactions over the phone**/*rooms in the hotel*?
The documents were **stored**/*sorted* by the intern in **filing cabinets**/*alphabetical order*.
He was **tired**/*tried* a long while after his **shift ended**/*arrest and indictment*.
My friend is **carving**/*craving* a massive **marble statue**/*cheeseburger and fries*.
She **fried**/*fired* every single one of the **chicken nuggets**/*employees and consultants*.
The **trial**/*trail* that was located in the **courthouse attracted media attention**/*woods attracted hikers and campers*.
Unlike the **males**/*meals* we discovered that the **females have dull feathers**/*transportation was not reimbursed*.
The chef's boss said his **bread**/*beard* definitely needs to be **baked at four hundred degrees**/*shaved off completely*.
These **genes**/*jeans* can be quickly and cheaply **sequenced by a lab**/*washed at the laundromat*.

C Computational Modeling Details

The noisy-channel model of Clark et al. (2025a,b) was used to generate surprisal values and posterior word error probabilities for all items in the study. The following parameters were used: `num_particles` = 128, `conditional_rejuv` = False, `second_pass_rejuv` = True, `second_pass_rejuv_p` = 1.0, `second_pass_rejuv_iters` = 3, `lm_method` = `gpt2`, `normal_alpha` = 3, `error_alpha` = 1, `prompt` = 1. A restricted vocabulary formed from the union of the top 5000 most frequent English words, according to the SUBTLEX-US dataset (Brysbaert and New, 2009) and all words appearing in the experimental materials was used. All other parameters were set at the default values.

Figure 4 shows the posterior distribution over actions at each word for an example sentence, extracted from the final set of 128 particles at the end of an inference run. Figure 5 shows the posterior distribution over inferred intended sentences for the same example sentence. The literal sentence is still the most likely candidate, but the alternative sentence with *licked* substituted for *kicked* is the next most likely candidate. Figure 6 shows the noisy-channel model's surprisal at each observation, with a comparison to baseline language model surprisal from GPT-2 (using the locally-constrained decoding method described in Clark et al. (2025b) to enforce a restricted vocabulary). Figure 7 shows the per-word acceptance rate for second-pass rejuvenations. Rejuvenations are most likely at words which have near neighbors that would make the sentence *a priori* more plausible.

We note that the "noisy-channel garden path" sentences provide a distinct challenge for computational models of language processing, and provide a direct contrast with incremental noisy-channel processing in sentences like *The storyteller could turn any story into an amusing antidote* (Ryskin et al., 2021), as modeled in (Clark et al., 2025b).

In those sentences, the anomalous word always occurred sentence-finally, and any error correction done by inferential comprehenders could be performed immediately upon observing the anomalous word. A key finding was that for such sentences, the surprisal on the final anomalous word would be lower under the noisy-channel model than under a baseline language model, due to the ability to recover an intended alternative (e.g. *anecdote*). Figure 6 show that no such reduction in incremental surprisal occurs at the anomalous predicate for this example, illustrating a "garden-path"-like effect. In contrast, rejuvenation moves for earlier choices enable the processor to arrive at inferences about intended alternatives.

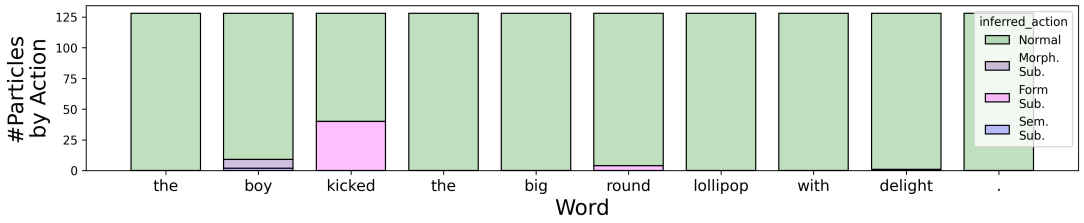


Figure 4: Posterior over actions for example sentence.

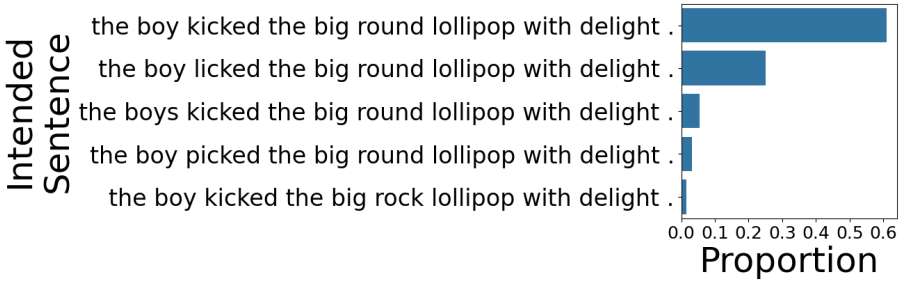


Figure 5: Posterior over inferred intended messages for example sentence.

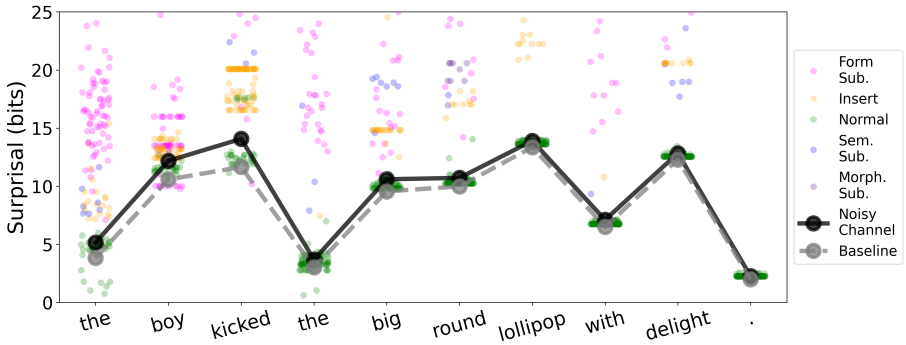


Figure 6: Per-word surprisal under baseline and noisy-channel language models.

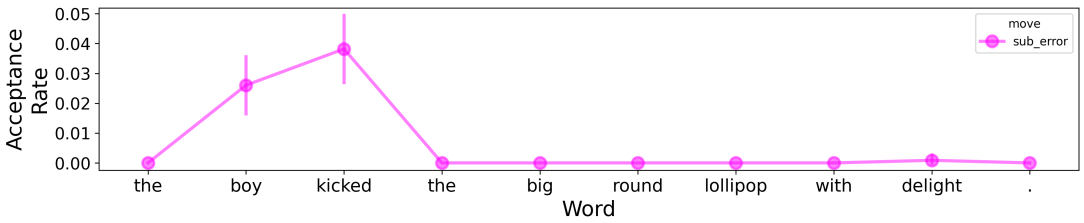


Figure 7: Rejuvenation acceptance rate at each word position for example sentence.

D Statistical Analysis: Full Results

Table 4: Fixed Effects Coefficients with 95% Credible Intervals (Regressions Out)

Coefficient	Estimate	95% CI
Intercept	**-3.883**	**[-4.219, -3.550]**
log_freq	**0.06**	**[0.035, 0.085]**
word_nchar	**0.08**	**[0.051, 0.109]**
surprisal	**0.055**	**[0.036, 0.073]**
word_num_in_sent	**-0.074**	**[-0.103, -0.046]**
pos_tagFUNCTION	-0.1	[-0.213, 0.012]
pos_tagADJ	-0.001	[-0.110, 0.110]
pos_tagVERB	0.048	[-0.047, 0.143]
pos_tagADV	-0.021	[-0.144, 0.101]
Preamble	-0.032	[-0.164, 0.102]
CriticalWord	0.073	[-0.058, 0.205]
Predicate	**0.309**	**[0.183, 0.435]**
Neighbor-GP	0.026	[-0.081, 0.133]
Typo	**0.125**	**[0.015, 0.231]**
Unrelated-GP	-0.011	[-0.116, 0.096]
Late-Error	0.055	[-0.049, 0.160]
Preamble × Neighbor-GP	0.045	[-0.114, 0.196]
CriticalWord × Neighbor-GP	-0.056	[-0.212, 0.106]
Predicate × Neighbor-GP	**0.333**	**[0.196, 0.469]**
Preamble × Typo	-0.054	[-0.214, 0.102]
CriticalWord × Typo	0.001	[-0.202, 0.198]
Predicate × Typo	**-0.45**	**[-0.597, -0.300]**
Preamble × Unrelated-GP	-0.013	[-0.176, 0.144]
CriticalWord × Unrelated-GP	0.059	[-0.094, 0.218]
Predicate × Unrelated-GP	**0.23**	**[0.091, 0.368]**
Preamble × Late-Error	-0.012	[-0.171, 0.145]
CriticalWord × Late-Error	0.024	[-0.135, 0.180]
Predicate × Late-Error	**0.166**	**[0.028, 0.305]**

Table 5: Fixed Effects Coefficients with 95% Credible Intervals (Regessions In)

Coefficient	Estimate	95% CI
Intercept	**-2.359**	**[-2.631, -2.083]**
log_freq	**-0.021**	**[-0.040, -0.003]**
word_nchar	**0.178**	**[0.159, 0.197]**
surprisal	**0.036**	**[0.025, 0.047]**
word_num_in_sent	**-0.134**	**[-0.158, -0.110]**
pos_tagFUNCTION	**-0.225**	**[-0.311, -0.140]**
pos_tagADJ	**0.103**	**[0.021, 0.183]**
pos_tagVERB	**0.096**	**[0.031, 0.163]**
pos_tagADV	**-0.138**	**[-0.235, -0.042]**
Preamble	**-0.199**	**[-0.313, -0.080]**
CriticalWord	**0.177**	**[0.069, 0.279]**
Predicate	**-0.187**	**[-0.294, -0.082]**
Neighbor-GP	**0.296**	**[0.218, 0.376]**
Typo	**-1.244**	**[-1.354, -1.136]**
Unrelated-GP	**0.389**	**[0.309, 0.468]**
Late-Error	**0.452**	**[0.376, 0.530]**
Preamble × Neighbor-GP	0.035	[-0.097, 0.165]
CriticalWord × Neighbor-GP	**0.369**	**[0.242, 0.499]**
Predicate × Neighbor-GP	**-0.134**	**[-0.258, -0.012]**
Preamble × Typo	-0.047	[-0.208, 0.116]
CriticalWord × Typo	0	[-0.202, 0.189]
Predicate × Typo	**-0.315**	**[-0.466, -0.161]**
Preamble × Unrelated-GP	0.085	[-0.043, 0.216]
CriticalWord × Unrelated-GP	**0.279**	**[0.156, 0.406]**
Predicate × Unrelated-GP	**-0.138**	**[-0.261, -0.018]**
Preamble × Late-Error	-0.117	[-0.242, 0.011]
CriticalWord × Late-Error	0.005	[-0.118, 0.127]
Predicate × Late-Error	-0.043	[-0.168, 0.080]

Table 6: Fixed Effects Coefficients with 95% Credible Intervals (Re-reading Time)

Coefficient	Estimate	95% CI
Intercept	**6.077**	**[6.002, 6.146]**
hu_Intercept	**0.726**	**[0.676, 0.775]**
log_freq	-0.011	[-0.039, 0.016]
word_nchar	**0.156**	**[0.138, 0.173]**
surprisal_nc	0.016	[-0.000, 0.033]
word_num_in_sent	-0.006	[-0.016, 0.004]
pos_tagFUNCTION	**0.043**	**[0.006, 0.081]**
pos_tagADJ	0.014	[-0.024, 0.051]
pos_tagVERB	0.021	[-0.012, 0.053]
pos_tagADV	-0.039	[-0.088, 0.010]
Preamble	**0.106**	**[0.038, 0.174]**
CriticalWord	0.074	[-0.007, 0.152]
Predicate	**0.093**	**[0.022, 0.160]**
Neighbor-GP	**-0.055**	**[-0.105, -0.003]**
Typo	**-0.148**	**[-0.226, -0.071]**
Unrelated-GP	-0.028	[-0.080, 0.022]
Late-Error	**-0.055**	**[-0.106, -0.004]**
Preamble × Neighbor-GP	-0.022	[-0.100, 0.053]
CriticalWord × Neighbor-GP	**0.215**	**[0.125, 0.310]**
Predicate × Neighbor-GP	-0.039	[-0.120, 0.043]
Preamble × Typo	0.088	[-0.018, 0.193]
CriticalWord × Typo	**0.456**	**[0.333, 0.578]**
Predicate × Typo	0.043	[-0.099, 0.184]
Preamble × Unrelated-GP	-0.065	[-0.142, 0.011]
CriticalWord × Unrelated-GP	**0.101**	**[0.007, 0.196]**
Predicate × Unrelated-GP	-0.071	[-0.150, 0.013]
Preamble × Late-Error	-0.015	[-0.096, 0.060]
CriticalWord × Late-Error	**0.133**	**[0.037, 0.232]**
Predicate × Late-Error	0.021	[-0.063, 0.104]
hu_log_freq	**0.078**	**[0.042, 0.114]**
hu_word_nchar	**-0.19**	**[-0.216, -0.164]**
hu_surprisal_nc	**-0.167**	**[-0.195, -0.141]**
hu_word_num_in_sent	**0.138**	**[0.132, 0.144]**
hu_pos_tagFUNCTION	**-0.067**	**[-0.130, -0.005]**
hu_pos_tagADJ	**-0.147**	**[-0.210, -0.085]**
hu_pos_tagVERB	**-0.168**	**[-0.221, -0.111]**
hu_pos_tagADV	0.025	[-0.055, 0.104]

E Pointwise Mutual Information

In a post-hoc anlaysis, we quantify the associative strength between a `CriticalWord` and its corresponding `Predicate` using pointwise mutual information (PMI) estimated from a masked language model, following past work (Wilcox et al., 2024b; Hoover et al., 2021). All estimates were obtained from RoBERTa-large (Liu et al., 2019), a bidirectional transformer pretrained on approximately 160 GB of English text.

PMI estimation

For a given variant "... [`CriticalWord`=c] ... [`Predicate`=p]" we estimate $\mathrm{PMI}(c\,;\,p)$ as

$$\log_2 \frac{P(c \mid p, \text{context})}{P(c \mid \text{context})}$$

using two masked-prediction queries to RoBERTa-large.

Numerator. The critical word position is replaced by `[MASK]` while the predicate phrase is left intact: For example, the sentence *The boy licked the big round lollipop with delight* would yield *The boy [MASK] the big round lollipop with delight*. The model's predicted probability at that mask position gives $P(c \mid p, \text{context})$.

Denominator. Both the critical word and the entire predicate phrase are replaced by mask tokens. The predicate phrase may span $k \geq 1$ subword tokens; it is replaced by exactly k consecutive `[MASK]` tokens (written `[MASK]`^{k}) to preserve sentence length: S_{den} = `Preamble [MASK] Intervening [MASK]`^{k}. The model's predicted probability at the critical-word mask gives $P(c \mid \text{context})$.

PMI score.

$$\begin{aligned}\mathrm{PMI}(c;\,p) \;&=\; \log_2 P(c \mid S_{\text{num}}) \\ &-\; \log_2 P(c \mid S_{\text{den}}),\end{aligned}$$

expressed in bits. A positive value indicates that the predicate raises the model's probability of the critical word above its baseline. Each item yields two `Plausible` variants and two `Neighbor-GP` variants, using a total of two unique `CriticalWord` values. Items for which either value of `CriticalWord` tokenizes to more than one subword token are excluded from this analysis. Only three items were excluded on this basis.

Per-item aggregate

Let $\pi^{(i)}_{jk} \equiv \mathrm{PMI}(c^{(i)}_j;\, p^{(i)}_k)$, where $j, k \in \{1, 2\}$ index the two variants of each item in the `Plausible` and `Neighbor-GP` conditions (e.g., *licked/kicked* and *ball into the net/lollipop with delight*). Item-level condition means are formed by averaging the two within-condition scores:

$$\widehat{\mathrm{PMI}}^{(i)}_{\texttt{Plausible}} \;=\; \tfrac{1}{2}\left[p^{(i)}_{11} + p^{(i)}_{22}\right],$$

$$\widehat{\mathrm{PMI}}^{(i)}_{\texttt{Neighbor-GP}} \;=\; \tfrac{1}{2}\left[p^{(i)}_{12} + p^{(i)}_{21}\right].$$

Statistical comparison uses a Wilcoxon signed-rank test on the per-item differences

$$\Delta^{(i)} \;=\; \widehat{\mathrm{PMI}}^{(i)}_{\texttt{Plausible}} - \widehat{\mathrm{PMI}}^{(i)}_{\texttt{Neighbor-GP}},$$

demonstrating that the two distributions of PMI values are almost certainly drawn from different distributions ($p < 0.001$).

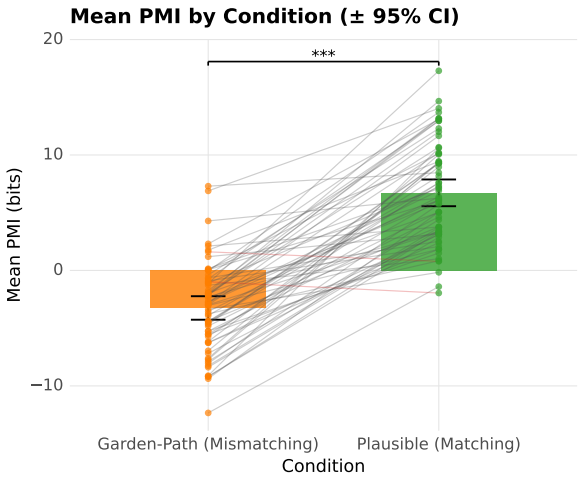


Figure 8: Pointwise mutual information across items and conditions (`Plausible` vs. `Neighbor-GP`) in the experimental materials. Lines connect values of $\pi^{(i)}_{11}$ and $\pi^{(i)}_{12}$, and connect values of $\pi^{(i)}_{22}$ and $\pi^{(i)}_{21}$. For all except two pairs, the `Plausible` condition has higher PMI than the `Neighbor-GP` condition.