A Model of Approximate and Incremental Noisy-Channel Language Processing

Thomas Hikaru Clark (thclark@mit.edu) **Jacob Hoover Vigly (jahoo@mit.edu)**

Edward Gibson (egibson@mit.edu) **Roger Levy (rplevy@mit.edu)**

MIT Department of Brain and Cognitive Sciences, 43 Vassar Street
Cambridge, MA 02139 USA

Abstract

How are comprehenders able to extract meaning from utterances in the presence of production errors? The noisy-channel theory provides an account grounded in Bayesian inference: comprehenders may interpret utterances non-literally in favor of an alternative with higher prior probability that is close under some error model. However, we lack implemented computational models of prior expectation and error likelihood capable of predicting human processing of arbitrary utterances. Here, we model sentence processing for "noisy" utterances as incremental and approximate probabilistic inference over intended sentences and production errors. We demonstrate that the model reproduces patterns in human behavior for anomalous sentences in three separate case studies from the noisy-channel literature. Our results offer a step towards an algorithmic account of inference during real-world language comprehension. Our model code, implemented in Gen, is available at `https://github.com/thomashikaru/noisy_channel_model`.

Keywords: Language processing; Bayesian inference; Probabilistic programming

Introduction

When studying the cognitive processes underlying language processing, it is common to assume that the input to a comprehender is a well-formed sentence corresponding to a speaker's intended meaning. Yet, language in everyday use frequently contains speech errors, typos, and other anomalies. How are comprehenders able to extract meaning from such "noisy" utterances, and where do the alternative interpretations for erroneous utterances come from? Understanding how humans are capable of recovering meaning from erroneous utterances can shed light on the mental computations involved in language processing in general.

The noisy-channel theory of language processing (Gibson et al., 2013; Levy, 2008) provides an account grounded in rational, Bayesian inference, where human inferences are influenced by both prior expectations and noise likelihood; comprehenders integrate information from an observed utterance with prior beliefs about intended sentences, using a generative model of how errors may occur. For example, comprehenders may implicitly assume that intended words may be skipped or substituted, or that unintended words may be inserted. Experimentally, comprehenders shown implausible sentences form non-literal interpretations in a significant portion of trials, with a preference for non-literal interpretations that involve more likely errors (Gibson et al., 2013). Additional experimental evidence shows that comprehenders' inferences are sensitive to the distribution of specific error categories in the environment and to supportive contexts (Chen et al., 2023; Ryskin et al., 2018). Meanwhile, the "good-enough processing" account emphasizes the role of shallow, heuristic-based processing during language comprehension to explain why comprehenders do not always notice errors, or revert to their priors when processing implausible sentences (Christianson, 2016; Ferreira & Patson, 2007; Li & Ettinger, 2023); other work has proposed a role for both rational inference and good-enough processing in how comprehenders interpret anomalous "illusion" sentences (Paape, 2024).

One limitation of existing accounts of language processing for atypical sentences is the general lack of implemented computational models capable of making rational inferences for arbitrary utterances. Large language models (Chang & Bergen, 2024; Radford et al., 2019; Vaswani et al., 2017) currently provide some of the best implemented models of language, and have been shown to produce results that resemble at least some human noisy-channel inferences (Cai et al., 2024; McCoy et al., 2023). Yet LLMs provide limited insight into the algorithms that comprehenders may employ during language processing, especially when recovering from errors (which may not be well-attested in training data), or for performing reanalysis of earlier material (something outside the scope of purely auto-regressive LLMs). Furthermore, a computational model of noisy-channel language comprehension should ideally account for how humans might perform complex inferences using limited cognitive resources; this may involve allocating more computation to challenging utterances than easy ones, something not reflected by LLMs (Hoover et al., 2023; Lieder & Griffiths, 2020).

In this work, we introduce a computational model of incremental noisy-channel language processing, which integrates next-word prediction with inferences about likely errors. We then report results from three case studies, demonstrating the model's ability to reproduce human inferences for anomalous sentences from the noisy-channel literature.

Computational Model

We design a computational model for noisy-channel language processing based on three key desiderata, in addition to the basic foundation of a Bayesian prior and likelihood. First, a model should be incremental: it should process sentences one word (or other linguistic unit) at a time, in the order that they

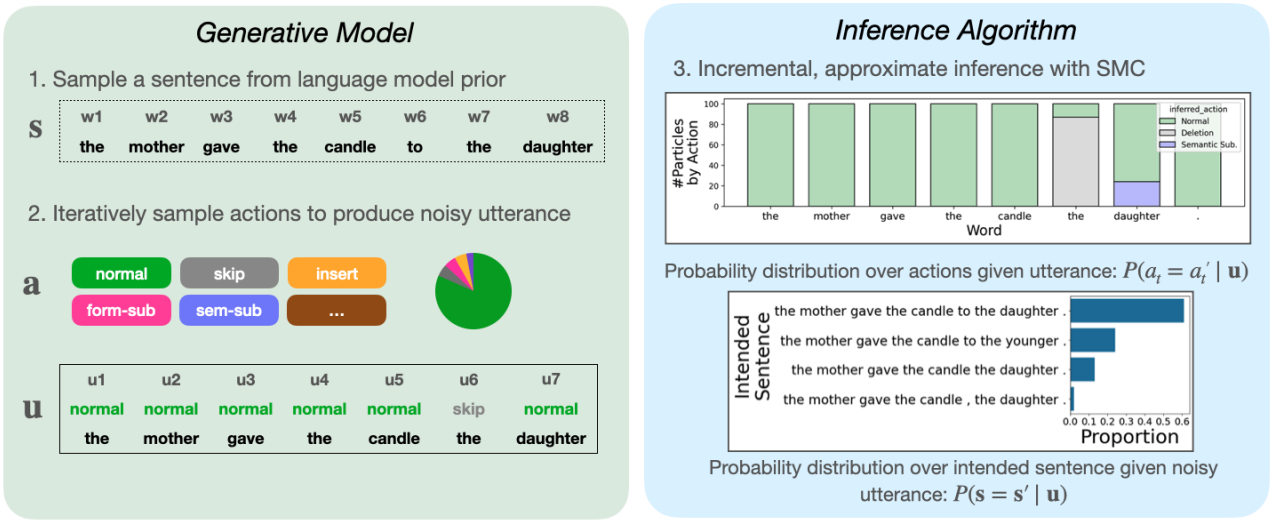

Figure 1: Model overview.

are observed (i.e. read or heard) by human comprehenders. Second, models should be approximate: performing exact inference over the vast space of alternative interpretations may be cognitively implausible, so a desirable model should be capable of approximating a desired posterior distribution using a variable amount of computation, e.g. via sampling-based methods (Hoover et al., 2023). Third, models should be capable of reanalysis: i.e., reinterpreting a previously processed word as an error in light of new observations.

Previous work has instantiated variations of a noisy-channel model that fulfill some but not all of these desiderata, in settings such as grammar correction using a weighted finite-state transducer (Park & Levy, 2011), modeling word recognition for children's speech by combining a contextual language model with a pronunciation model (Meylan et al., 2023), or modeling ERPs for anomalous sentences by weighing a literal utterance against a given alternative (Li & Ettinger, 2023). Other relevant work has instantiated noisy-channel principles computationally, but by modeling noise in memory, rather than in production (Futrell et al., 2020; Hahn et al., 2022).

Here, we extend previous work by building a general and modular noisy-channel model that performs incremental and approximate inference for possibly anomalous sentences. Our model consists of three parts: a) a language model, b) an error model, and c) an inference algorithm; the language model and error model collectively form a generative model of language production, and the inference algorithm is deployed on observed utterances to infer latent variables such as the intended sentence and error identities (see overview in Figure 1). We then compare model outputs to established human behavioral results. The model is implemented in the Gen probabilistic programming language (Cusumano-Towner et al., 2019), which automates many of the mathematical computations involved in performing inference. Importantly, our model does not require training on noisy sentences, and its parameters are either chosen based on prior knowledge or sampled from wide priors and inferred jointly during inference.

Language Model

A language model (LM) is a statistical model of language that assigns probabilities to sequences of linguistic units, such as words. Sampling a sentence proceeds by iteratively sampling words from the LM. In our framework, the LM serves simply as a prior over intended sentences, thus we can use a relatively small LM without specialized mechanisms aimed at eliciting reasoning-like behavior (Wei et al., 2023). We use GPT-2 (Radford et al., 2019), which has been shown to encode predictability in a way that correlates strongly with human reading times (Shain et al., 2024).[1] GPT-2 uses a vocabulary of byte-pair-encoding tokens, while we model a probability distribution over words; we thus adapt GPT-2 using token masking to a restricted vocabulary $\mathcal{V}$ of the 5000 most frequent English words according to the SUBTLEX-US dataset (Brysbaert & New, 2009).[2]

Error Model

Given a sentence $\mathbf{s}$ sampled from the language model, the error model iterates word-by-word through $\mathbf{s}$, and at each time step t samples an action $\mathbf{a}_t$ independently from a probability distribution over 6 action types: **normal** production, **insertion**, **skip**, and **form-based**, **semantic**, and **morphological** substitutions. This probability distribution over actions, which is a latent variable shared by the entire sentence, is drawn from a Dirichlet prior with concentration parameter 10 for **normal** and 1 for all error types. Because of insertions and deletions, the index of the current intended word within

[1] We use the `hfppl` library (Lew, Zhi-Xuan, et al., 2023) for language model caching to speed up inference.

[2] For each experiment, we take the union of this base vocabulary with all word types in the experimental dataset.

$\mathbf{s}$ may not be equal to t; we use the notation $\text{idx}(t)$ to denote the index in $\mathbf{s}$ that should be produced at time t.

At time t, given $\mathbf{s}_{\text{idx}(t)}$ and $\mathbf{a}_t$, the error model generates the output word by applying symbolic rules. For the **normal** action, the output word will simply be $\mathbf{s}_{\text{idx}(t)}$ itself. For **skip**, the output word will be $\mathbf{s}_{\text{idx}(t)+1}$. For **form-based** substitutions, the output word is sampled from a probability distribution over $\mathcal{V}$ where each word's probability is monotonic decreasing in its Levenshtein distance, denoted $\text{Lev}(\cdot,\cdot)$, from $\mathbf{s}_{\text{idx}(t)}$ (Levenshtein, 1965): $p(a \mid b) \propto \beta_1^{\text{Lev(a, b)}}$, where $\beta_1 \sim \text{Beta}(2, 11)$ is a latent variable quantifying how peaked or flat the distribution is, and where $p(a \mid b)$ is clamped to 0 for pairs where $\text{Lev}(a,b) > 5$. For **semantic** substitutions, the output word is sampled from a probability distribution over $\mathcal{V}$ where each token's probability is monotonically decreasing in its cosine distance from $\mathbf{s}_{\text{idx}(t)}$ in the `GloVe` semantic embedding space (Pennington et al., 2014): $p(a \mid b) \propto \text{cosineSim(a, b)}^{\beta_2}$, where $\beta_2 \sim \text{Gamma}(6, 1)$ is another latent variable governing the distribution's peakedness, and where $p(a \mid b)$ is clamped to 0 for items outside the 20 closest neighbors. For insertions, the output word is sampled randomly from the unigram frequency distribution over $\mathcal{V}$, independently of context. For **morphological** substitutions, we replace $\mathbf{s}_{\text{idx}(t)}$ with another form of the same lemma if one exists in $\mathcal{V}$, e.g. *kick* $\rightarrow$ *kicks*. We denote the sequence of output words as $\mathbf{u}$.

This error model is not intended to be an accurate model of language production, but rather a proxy for a rational comprehender's intuitive theory of noisy production. While it leaves out some purported basic error operations such as exchanges (Poppels & Levy, 2016), it can generate a wide range of plausible transformations of a given intended sentence. Some sources of uncertainty are encoded as latent variables and included in the inference problem (e.g., β_1 and β_2). Other model choices, such as the use of GloVe embeddings or the concentration parameters for the Dirichlet prior, are fixed properties of the model. We leave further exploration of the space of error models to future work.

Inference Algorithm: Particle Filter

Given an observed utterance, we perform inference on the latent variables in the generative model. We use a particle filter, a standard Sequential Monte Carlo algorithm (Naesseth et al., 2024). Particle filters, while originally introduced for inference problems unrelated to language processing, have been used to model the effect of limited memory on human linguistic inferences (Levy et al., 2008).

The particle filter maintains a set of K, particles (for all results reported in this paper, we set $K = 64$), each corresponding to a hypothesis about the model state, i.e. the values of all latent random variables in the generative model up to the current time step. Each particle $x_t^{(i)}$ is associated with a weight $w_t^{(i)}$, which, when normalized across particles, serves as an approximation to the probability of the particle's state given the observations (Chopin & Papaspiliopoulos, 2020). We use the particle filter to infer the posterior distribution over states, given a set of observations: $p(x_t \mid \mathbf{u}_{1:t})$. At time t, the particle filter incrementally samples a new extended state for each particle, which expresses a hypothesis about $\mathbf{s}_{\text{idx}(t)}$ and $\mathbf{a}_t$. In principle, each particle can now be scored in terms of how well it explains the new observation $\mathbf{u}_t$.

However, due to the symbolic rules in the error model, new particle states randomly sampled from the generative model are likely to be incompatible with the observation (e.g. $\mathbf{s}_{\text{idx}(t)} = \textit{mother}, \mathbf{a}_t = \textbf{normal}, \mathbf{u}_t = \textit{boy}$). We thus use a custom proposal function $q(\cdot)$, which assigns $\mathbf{s}_{\text{idx}(t)}$ heuristically, by either setting it equal to $\mathbf{u}_t$, sampling a form-based or semantic neighbor of $\mathbf{u}_t$, or sampling from the LM-induced next-word distribution given the context $\mathbf{u}_{1:t-1}$. It then samples an action with non-zero probability of generating $\mathbf{u}_t$. We then apply an importance weight correction in the weight update to offset the bias introduced by this proposal function.[3] Particles are resampled at each time step, which resets their weights to a uniform distribution.

Model Surprisal and Model Inferences. We define particle filter surprisal as the negative log of the mean unnormalized particle weight, which approximates the conditional probability of an observation in context: $p(\mathbf{u}_t \mid \mathbf{u}_{1:t-1}) = \int p(\mathbf{u}_t \mid x_t) p(x_t \mid \mathbf{u}_{1:t-1})\, dx_t \approx \frac{1}{K} \sum_{i=1}^{K} w_t^{(i)}$ Intuitively, surprisal is low when the current observation is explainable either as a high-probability continuation in normal production, or an error that is likely under the error model. When inference reaches the end of the utterance, each particle contains a specific hypothesis about the intended sentence $\mathbf{s}$ and the sequence of actions that map from $\mathbf{s}$ to $\mathbf{u}$, thus we can extract a posterior distribution over $\mathbf{s}$ and each $\mathbf{a}_t$ by simply drawing samples from the final particle filter state.

Rejuvenation. While incremental processing is the default in our model, we also employ particle *rejuvenation* to increase particle diversity. Rejuvenation refers to modifying the random choices of a particle from an earlier time step, and is a popular addition to Sequential Monte Carlo inference (Gilks & Berzuini, 2001; Lew, Matheos, et al., 2023; Lew, Zhi-Xuan, et al., 2023). Without rejuvenation, earlier random choices are never revised; this is problematic given a finite particles set, since globally promising particles may be filtered out in favor of locally high-scoring ones. We also speculate that there is a cognitive significance to rejuvenation in the context of rational inference models — rejuvenation allows inference to proceed with fewer particles, at the cost of possible extra computation later during reanalysis.

A specific rejuvenation proposal function modifies the choices made for some of the random variables in a particle x_t, yielding a modified particle x_t'. In this work, we employ two rejuvenation proposals: one which proposes that some earlier word was a substitution error, and one which proposes that some earlier word was either erroneously inserted or deleted. Crucially, these proposals are constructed

[3] The new weight $w_t^{(i)}$ for particle $x_t^{(i)}$ at time t is given by: $w_t^{(i)} = p(\mathbf{u}_t \mid x_t^{(i)}) \frac{p(x_t^{(i)} \mid x_{t-1}^{(i)})}{q(x_t^{(i)} \mid x_{t-1}^{(i)}, \mathbf{u}_t)}$, which is calculated automatically in Gen.

Condition	Sentence
PO-plaus.	The boy threw the apple to the girl.
DO-plaus.	The boy threw the girl the apple.
PO-implaus.	The boy threw the girl to the apple.
DO-implaus.	The boy threw the apple the girl.
T-plaus.	The earthquake shattered the house.
P-plaus.	The house shattered from the earthquake.
T-implaus.	The house shattered the earthquake.
P-implaus.	The earthquake shattered from the house.

Table 1: Example sentences from the Gibson et al. (2013) study.

Condition	Sentence
SSS	The cause of the problem was investigated.
SPS	The cause of the problems was investigated.
SSP	The cause of the problem were investigated.
SPP	The cause of the problems were investigated.

Table 2: Example sentences from the Qian and Levy (2023) study. The condition codes denote whether the subject, intervening noun, and verb are singular 'S' or plural 'P', respectively.

to be involutive — applying a proposal $g(\cdot)$ to a particle twice has some non-zero probability of leaving it unchanged (Cusumano-Towner et al., 2020; Neklyudov et al., 2020). This property allows us to employ the Metropolis-Hastings algorithm to probabilistically accept $x_t{}'$ or keep x_t. We include rejuvenation moves both during incremental inference (targeting earlier words within a 3-word lookback window at each time step) and after incremental processing is complete (targeting all words in the sentence).

To consider why rejuvenation may be relevant for noisy-channel processing, consider a case where an error is not apparent until late in the sentence, such as '*The fires were quickly pumped up by the mechanic*'. A particle which maps the observation '*fires*' to the intended word '*tires*' via a substitution would have a high score in light of the full sentence. Yet without a large particle count, such a particle is unlikely to survive the resampling steps that occur before the full sentence is seen. With rejuvenation, such particles can be proposed and accepted later even if they were filtered out or were never sampled to begin with.

Experiments

Experiment 1: (Gibson et al., 2013)

Experimentally, it has been shown that readers often interpret an implausible sentence non-literally, as if it were a more plausible sentence differing by a small edit (Table 1) (Gibson et al., 2013). Furthermore, there is an asymmetry between insertion and deletion errors: comprehenders were more likely to form a non-literal interpretation consistent with a deletion error, compared to an insertion error, consistent with the Bayesian "size principle" (Xu & Tenenbaum, 2007), which holds that any particular insertion, chosen from the entire vocabulary, is less likely than the deletion of one word in the sentence.

Here we ask whether the particle filter model can provide an algorithmic account of non-literal interpretations for implausible utterances and the deletion-insertion asymmetry. We take the dative alternation and transitive alternation materials from Gibson et al. (2013) and run posterior inference with our model on each item individually. The model returns a distribution over inferred actions for each word in the sentence; intuitively, a high posterior probability assigned to the **normal** action at a critical word implies a "literal" interpretation. We define the 'critical word' as the first word of the final phrase in the sentence, i.e. the point at which an error could be hypothesized in the implausible materials. If the rational model is humanlike, it should make more non-literal interpretations for implausible sentences than for plausible sentences, and for the DO and transitive structures than for PO or Preposition structures.

Experiment 2: Qian and Levy (2023)

A second test case of the noisy-channel theory comes from ambiguous errors – when there are multiple possible ways to correct an error in a sentence, how do comprehenders decide how to edit (Qian & Levy, 2023)? In sentences with agreement errors (Table 2), human edits were sensitive both to the prior probability of sentences and to the statistics of errors in language production data; this contrasts with a naive model in which people make edits based on part of speech or linear order but without reference to a generative model of errors. Here, we test if our model can recover the pattern of human responses in this dataset. We run our model on the experimental sentences with singular subjects and plural verbs. At each of the subject and verb, we compute the posterior probability that the word is an error (i.e., a non-**normal** action). We define the model verb-edit probability as $P(a_{\textbf{verb}} = \text{error} \mid \mathbf{u})/(P(a_{\textbf{verb}} = \text{error} \mid \mathbf{u}) + P(a_{\textbf{subject}} = \text{error} \mid \mathbf{u}))$, and test whether this quantity predicts variance in human verb-edit preferences at the item level.

Experiment 3: Ryskin et al. (2021)

We further test the model on the materials of Ryskin et al. (2021), which include grammatical errors, substitution errors, and unrelated continuations in the same contexts, for example: "*The storyteller could turn any incident into an amusing* {*anecdote/anecdotes/antidote/hearse*}". We label these four conditions as Normal, Ungrammatical, Neighbor, and Unrelated, respectively. In an EEG study, words that were explainable as errors (i.e., Ungrammatical and Neighbor) were shown to follow a distinct profile from Normal and Unrelated words. If our model captures human behavior well, we should see a dissociation between the four conditions in the experiment. Specifically, we expect to see a high rate of literal

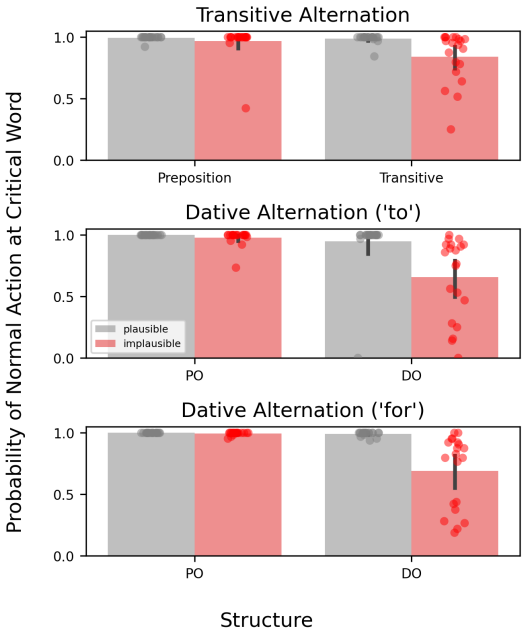

Figure 2: Model inferences for the dative and transitive alternations in (Gibson et al., 2013).

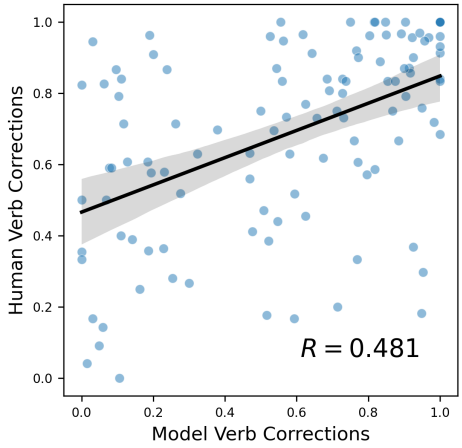

Figure 3: Model verb edit preferences plotted against human verb edit preferences for ambiguous agreement errors in (Qian & Levy, 2023). The solid line denotes the best linear fit with 95% confidence intervals.

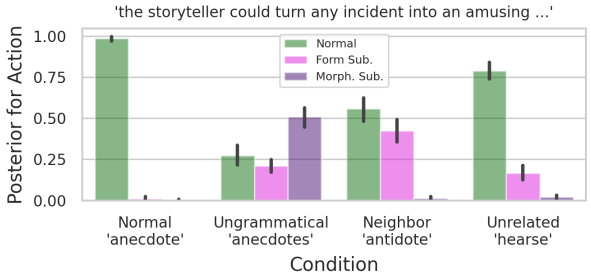

Figure 4: Average posteriors over actions at the critical word, by condition and action in (Ryskin et al., 2018). Error bars denote 95% CIs.

interpretations for the Normal and Unrelated conditions, and a high rate of inferential interpretations for the Neighbor and Ungrammatical conditions (corresponding to error-correction and recovery of an alternative interpretation). Turning to surprisal, we expect to see the lowest surprisal for Normal and the highest for Unrelated, with intermediate values for Neighbor and Ungrammatical. Crucially, surprisal from the noisy-channel model for the Neighbor and Ungrammatical conditions should be lower than for the baseline, non-noisy-channel model, since the error can be explained in terms of the error model.

Results

Experiment 1

Using our noisy-channel particle filter model, we find a tendency for the same qualitative pattern of inferences as seen in humans (Figure 2). Our proxy for literal interpretation was close to 1.0 for plausible sentences, while it was significantly lower for implausible sentences. Non-literal interpretations in the implausible sentences were largely driven by the DO-implausible and Transitive-implausible sentences, i.e. those consistent with a single deletion error applied to a plausible counterpart, thus showing an insertion-deletion asymmetry.

Experiment 2

Figure 3 shows the relationship between model verb-edit preference and the human verb-edit preference, across items. Model edit preference is positively correlated with human edit preferences ($R = 0.481$). However, we note that this degree of correlation is lower than the average correlation of split halves of the human data ($R = 0.825$), showing that the model does not fully explain the pattern of human results. These results were achieved without fitting to any human data and with uninformative priors about the relative probabilities of different types of errors; future work can consider data-driven methods for increasing model alignment with human performance.

Experiment 3

Figure 4 shows the average posterior over actions in each of the four conditions, averaged across experimental items. First, we observe the highest rates of literal interpretations (quantified using the posterior over action at the critical word) in the Normal and Unrelated conditions, and higher rates of non-literal interpretations in the Ungrammatical and Neighbor conditions. Next, noisy-channel surprisal is lower than baseline surprisal in those conditions where the error is explainable in terms of the error model (Ungrammatical and Neighbor), and slightly higher in the other conditions (Figure 5).

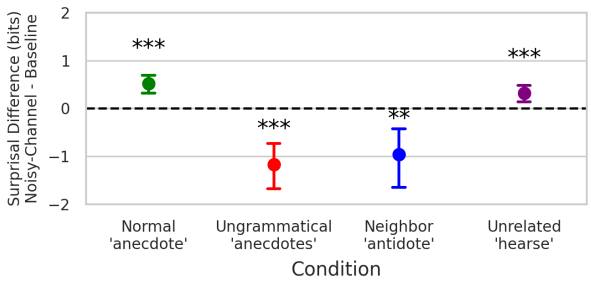

Figure 5: Difference in surprisal between the noisy-channel and baseline models at the critical word across conditions. Stars indicate standard p-values using a paired t-test comparing items within a condition. Error bars denote 95% CIs.

Discussion

Various studies have shown that people make inferences consistent with rational inference during comprehension of anomalous sentences. However, it has remained unclear how, at an algorithmic level, comprehenders may be solving this problem, especially given constraints on human cognitive resources and the large search space of alternative interpretations. In this work, we put forward a candidate algorithmic account of noisy-channel processing, made up of interpretable and modular building blocks. The LM module encodes a preference for interpretations corresponding to high-probability messages, while the error model is used to assess the probability of particular errors. An approximate and incremental inference algorithm attempts to balance the strengths of rational computational models with the constraints on cognitive resources.

Through three case studies, we show that our rational model captures qualitative and quantitative patterns of human noisy-channel inference reported in the literature, spanning sentences that can be interpreted as having insertion, deletion, substitution, and morphological errors (Gibson et al., 2013; Qian & Levy, 2023; Ryskin et al., 2021). In Experiment 1, the Bayesian size principle falls out naturally from our generative model – while insertions and deletions have the same prior probability, because the space of possible insertions is larger, any particular insertion is less likely. Experiment 3 showed a match to human editing preferences, though the rational model still falls short of inter-participant reliability, raising the question of what other cues people may use to inform their edits. In Experiment 3, the patterns of surprisal in from the model indicate how incrementally, a rational comprehender may be able to use context to reconstruct an intended word when it is a neighbor (under some error model) of a plausible continuation. However, prior work has also shown that incremental surprisal alone does not fully account for processing difficulty in cases that require reanalysis (van Schijndel & Linzen, 2021; van Schijndel & Linzen, 2018); additional inference mechanisms, such as rejuvenation, may provide a new way to model such forms of linguistic repair.

Future work can investigate the relationship between task performance and computational resources. Varying the number of particles used during inference could lead to qualitatively different inferences, e.g. plausible alternative interpretations that require more computation to find; this variation can be compared to human comprehenders placed under varying degrees of cognitive load. We can also model online belief updates regarding the prevalence of different types of errors, as in Ryskin et al. (2018), who showed that comprehenders' inferences are sensitive to the rate and type of errors appearing in experimental materials. This can be modeled by performing inference on batches of utterances rather than single utterances, allowing for greater accumulation of evidence about the underlying probabilities of different error categories.

Future work can also develop linking hypotheses between the inference algorithm and more fine-grained human data, such as reading times (Bicknell et al., 2020; Frazier & Rayner, 1982; Rayner, 1998) and event-related potentials (ERPs) (Li & Ettinger, 2023; Li & Futrell, 2024). Several recent works have considered the role of prediction error and reasoning about alternatives in comprehension, as measured by reading times (Giulianelli et al., 2025; Meister et al., 2024). This work complements such work by a) considering anomalous and erroneous linguistic input, not just prediction errors; b) explicitly instantiating an error model built on symbolic operations such as substitutions, insertions, and skips; and c) modeling both incremental and reanalytical behavior. To this end, novel noisy-channel datasets can be designed specifically so that anomalies are difficult to resolve without reanalysis, and reading paradigms that allow for regressive reading, such as eye-tracking or mouse-tracking, can be employed rather than the Self-Paced Reading or Maze paradigms used in many reading time datasets (Wilcox, Ding, et al., 2024; Wilcox, Pimentel, et al., 2024). Human evaluation of reading behavior, paired with model evaluations comparing a variety of rejuvenation strategies, can shed light on the algorithms used by humans during robust real-world comprehension.

Conclusion

We introduce a model of incremental and approximate noisy-channel inference, and evaluate it in several case studies. The model recapitulates attested patterns of non-literal interpretations in humans, indicating that human-like patterns emerge when combining next-word prediction with rational inference over possible errors. These results provide a step towards an interpretable, algorithmic account of robust and resource-rational language processing. Our modular framework also provides a way to explore the space of error models and inference strategies and their role in noisy-channel comprehension.

Acknowledgments

We would like to thank members of the following labs for helpful comments: TedLab, Computational Psycholinguistics lab, CoCoSci, the Probabilistic Computing Project, and the UC Irvine Language Processing Group, as well as audiences at the Human Sentence Processing Conference, for helpful comments. This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. SMA-2404644.

References

Bicknell, K., Levy, R., & Rayner, K. (2020). Ongoing Cognitive Processing Influences Precise Eye-Movement Targets in Reading. *Psychological Science*, *31*(4), 351–362.

Brysbaert, M., & New, B. (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.

Cai, Z. G., Duan, X., Haslett, D. A., Wang, S., & Pickering, M. J. (2024, March). Do large language models resemble humans in language use?

Chang, T. A., & Bergen, B. K. (2024). Language Model Behavior: A Comprehensive Survey. *Computational Linguistics*, 1–58.

Chen, S., Nathaniel, S., Ryskin, R., & Gibson, E. (2023). The effect of context on noisy-channel sentence comprehension. *Cognition*, *238*, 105503.

Chopin, N., & Papaspiliopoulos, O. (2020). Particle Filtering. In N. Chopin & O. Papaspiliopoulos (Eds.), *An Introduction to Sequential Monte Carlo* (pp. 129–165). Springer International Publishing.

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 817–828.

Cusumano-Towner, M. F., Lew, A. K., & Mansinghka, V. K. (2020, July). Automating Involutive MCMC using Probabilistic and Differentiable Programming [arXiv:2007.09871 [stat]].

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., & Mansinghka, V. K. (2019). Gen: A General-purpose Probabilistic Programming System with Programmable Inference. *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 221–236.

Ferreira, F., & Patson, N. D. (2007). The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

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

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, *44*(3), e12814.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Gilks, W. R., & Berzuini, C. (2001). Following a moving target—Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(1), 127–146.

Giulianelli, M., Wallbridge, S., Cotterell, R., & Fernández, R. (2025, February). Incremental Alternative Sampling as a Lens into the Temporal and Representational Resolution of Linguistic Prediction.

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), e2122602119.

Hoover, J. L., Sonderegger, M., Piantadosi, S. T., & O'Donnell, T. J. (2023). The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing. *Open Mind*, *7*, 350–391.

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*.

Levy, R. (2008). A Noisy-Channel Model of Human Sentence Comprehension under Uncertain Input. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243.

Levy, R., Reali, F., & Griffiths, T. L. (2008). Modeling the effects of memory on human online sentence processing with particle filters.

Lew, A. K., Matheos, G., Zhi-Xuan, T., Ghavamizadeh, M., Gothoskar, N., Russell, S., & Mansinghka, V. K. (2023). SMCP3: Sequential Monte Carlo with Probabilistic Program Proposals [ISSN: 2640-3498]. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 7061–7088.

Lew, A. K., Zhi-Xuan, T., Grand, G., & Mansinghka, V. K. (2023, November). Sequential Monte Carlo Steering of Large Language Models using Probabilistic Programs.

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

Li, J., & Futrell, R. (2024, May). An information-theoretic model of shallow and deep language comprehension [arXiv:2405.08223 [cs, math]].

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023, September). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve [arXiv:2309.13638 [cs]].

Meister, C., Giulianelli, M., & Pimentel, T. (2024, October). Towards a Similarity-adjusted Surprisal Theory [arXiv:2410.17676 [cs]].

Meylan, S. C., Foushee, R., Wong, N. H., Bergelson, E., & Levy, R. P. (2023). How adults understand what young children say [Publisher: Nature Publishing Group]. *Nature Human Behaviour*, *7*(12), 2111–2125.

Naesseth, C. A., Lindsten, F., & Schön, T. B. (2024, December). Elements of Sequential Monte Carlo [arXiv:1903.04797 [stat]].

Neklyudov, K., Welling, M., Egorov, E., & Vetrov, D. (2020). Involutive MCMC: A unifying framework. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 7273–7282.

Paape, D. (2024). How do linguistic illusions arise? Rational inference and good-enough processing as competing latent processes within individuals. *Language, Cognition and Neuroscience*, *39*(10), 1334–1365.

Park, Y. A., & Levy, R. (2011, June). Automated Whole Sentence Grammar Correction Using a Noisy Channel Model. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 934–944). Association for Computational Linguistics.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Poppels, T., & Levy, R. P. (2016). Structure-sensitive Noise Inference: Comprehenders Expect Exchange Errors.

Qian, P., & Levy, R. P. (2023, September). Comprehenders' Error Correction Mechanisms are Finely Calibrated to Language Production Statistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners, 24.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *124*(3), 372–422.

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, *181*, 141–150.

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, *158*, 107855.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121.

van Schijndel, M., & Linzen, T. (2021). Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, *45*(6), e12988.

van Schijndel, M., & Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. *Neural Network Models*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023, January). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [arXiv:2201.11903 [cs]].

Wilcox, E. G., Ding, C., Sachan, M., & Jäger, L. A. (2024). Mouse Tracking for Reading (MoTR): A new naturalistic incremental processing measurement tool. *Journal of Memory and Language*, *138*, 104534.

Wilcox, E. G., Pimentel, T., Meister, C., & Cotterell, R. (2024). An information-theoretic analysis of targeted regressions during reading. *Cognition*, *249*, 105765.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.