UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

An information-theoretic model of shallow and deep language comprehension

Permalink

https://escholarship.org/uc/item/1fd682nd

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Li, Jiaxuan

Futrell, Richard

Publication Date

2024

Peer reviewed

An information-theoretic model of shallow and deep language comprehension

Jiaxuan Li (jiaxuan.li@uci.edu)
Department of Language Science, University of California Irvine
Irvine, CA 92617 USA

Richard Futrell (rfutrell@uci.edu)
Department of Language Science, University of California Irvine
Irvine, CA 92617 USA

Abstract

A large body of work in psycholinguistics has focused on the idea that online language comprehension can be shallow or 'good enough': given constraints on time or available computation, comprehenders may form interpretations of their input that are plausible but inaccurate. However, this idea has not yet been linked with formal theories of computation under resource constraints. Here we use information theory to formulate a model of language comprehension as an optimal tradeoff between accuracy and processing depth, formalized as bits of information extracted from the input, which increases with processing time. The model provides a measure of processing effort as the change in processing depth, which we link to EEG signals and reading times. We validate our theory against a large-scale dataset of garden path sentence reading times, and EEG experiments featuring N400, P600 and biphasic ERP effects. By quantifying the timecourse of language processing as it proceeds from shallow to deep, our model provides a unified framework to explain behavioral and neural signatures of language comprehension.

Keywords: Good-enough comprehension; Information theory; Rate–Distortion Theory; N400; P600; Reading Time; Syntactic Ambiguity

Introduction

Language comprehension can be inaccurate in the sense that a comprehender's interpretations of their input may not reflect the input veridically. For example, given input such as

(1) The storyteller could turn any incident into an amusing *antidote* ...

a comprehender may initially interpret the anomalous word *antidote* as the more plausible *anecdote*, and the misinterpretation may persist (Barton & Sanford, 1993; Erickson & Mattson, 1981; Ferreira, 2003; Ferreira et al., 2002; Fillenbaum, 1971, 1974; Sanford, 2002). Similarly, processing of syntactic structure can be inaccurate, with comprehenders interpreting a sentence such as

(2) When the little girl attacked the lamb remained calm.

as meaning that the little girl attacked the lamb, while the 'correct' interpretation derived through a full parse of the sentence is that the girl was attacked by the lamb (Christianson et al., 2001; Ferreira et al., 2001).

Such observations have motivated the idea that language processing can be 'good enough' or heuristic: comprehenders are willing to tolerate some level of inaccuracy in their interpretations in return for not having to expend too much computational effort (Ferreira & Patson, 2007; Ferreira et al., 2002). More generally, language processing has been held to proceed through stages reflecting different levels of depth, with an initial shallow, heuristic, and local stage of processing, which may be followed by deeper, global, and more effortful processing when necessary (Bever, 1970; Frazier & Fodor, 1978; Frazier & Rayner, 1982; Gouvea et al., 2010; Hoeks et al., 2004; MacDonald et al., 1994; Tabor et al., 2004; Trueswell et al., 1993; Van Gompel et al., 2000).

While the idea of shallow processing explains a wide variety of phenomena and has intuitive appeal, the idea of 'depth of processing' has yet to be made precise in a way that links it formally to theories of computation under resource constraints (Lewis et al., 2014; Lieder & Griffiths, 2020). We propose a computational model in which language comprehension instantiates optimal tradeoffs between accuracy and processing depth, where processing depth is measured information-theoretically in terms of bits of information extracted from perceptual input. Our model joins recent explanations of human and animal perception and behavior based on rate–distortion theory (Arumugam et al., 2022; Bhui et al., 2021; Cheyette & Piantadosi, 2020; Futrell, 2023; Mollica et al., 2021; Sims, 2016, 2018; Zaslavsky et al., 2018, 2020). Furthermore, our model provides a link between notions of depth of processing, processing effort, and processing time: we hold that processing depth increases with time, and that an increase in processing depth is reflected in processing effort, which in turn is reflected in EEG signals and reading times.

Below, we present our model using comprehension of Example (1) as a running example. Then we validate the model by simulating N400 and P600 effects elicited by a variety of complex semantic and syntactic manipulations from two experiments, and by simulating reading times for garden-path constructions in the Syntactic Ambiguity Processing dataset (Huang et al., 2023), successfully predicting the relative magnitudes of garden path slowdowns in reading time.

Model

We characterize online language comprehension using an **interpretation policy**, which represents a comprehender's probability distribution on interpretations w given perceptual inputs x. We hold that comprehenders' interpretation policies minimize a **distortion metric**—a measure $d(w, x)$ of how bad it is to form an interpretation w given input x—subject to
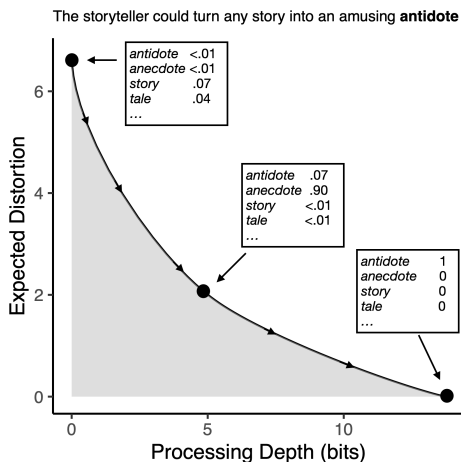
Figure 1: Tradeoff between distortion and processing depth (KL divergence) in optimal interpretation policies for the given input. Each location in the white part of the plane represents a possible interpretation policy for the input; the tradeoffs in the gray region are unachievable. The black line shows the *efficient frontier* of policies that achieve the minimal distortion for a given level of processing depth. We hold that interpretation policies move down this frontier with increasing processing time.

a constraint on bits of information extracted from the sensory input. An example of interpretation policies and their tradeoffs is shown in Figure 1. We furthermore posit that the tradeoff of distortion and processing depth depends on time: while a comprehender is perceiving input x, more information becomes available to them, and so their interpretation policy moves along the efficient frontier in Figure 1, evolving from a shallow, inaccurate policy to a a deep, accurate policy while maintaining an optimal tradeoff.

Formal model We propose that the comprehenders' interpretation policy $p_t(w|x)$ at a given time t is chosen to minimize the expected distortion subject to a constraint on the depth of processing, $D(t)$:

$$\underset{p_t(w|x)}{\text{minimize}} \underset{p(x)p_t(w|x)}{\mathbb{E}} [d(w,x)] \text{ subject to } D(t) \leq C(t), \quad (1)$$

where the processing depth $D(t)$ is the average KL divergence (Cover & Thomas, 2006) between the interpretations p_t and the 'default' policy $p_0(w)$ representing a comprehender's prior expectations:

$$D(t) = \underset{p(x)}{\mathbb{E}} \left[D_{\text{KL}} \left[p_t(w \mid x) \| p_0(w) \right] \right], \quad (2)$$

and $C(t)$ is a constraint on the processing depth achievable at time t (Cheyette & Piantadosi, 2020). The KL divergence reflects the extent to which the interpretation policy $p_t(w \mid x)$ moves away from the default interpretations $p_0(w)$ in response to the information provided by the input x.
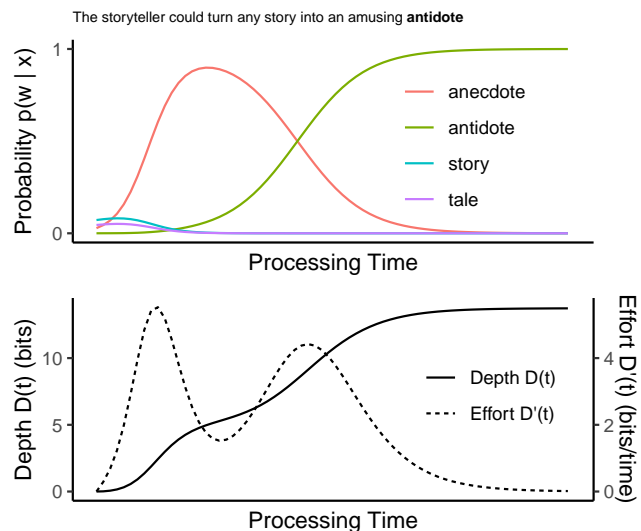

Figure 2: **Top.** Probabilities of four interpretations w given input $x =$ "story" in the given context as a function of processing time, as predicted from Eq. 4 with parameters described in the text. **Bottom.** Measures of processing effort over time for the same input.

To summarize, we are optimizing the processing accuracy (minimizing distortion) in Eq. 1 with a processing depth constraint as defined in Eq. 2. Applying the method of Lagrange multipliers, the optimal interpretation policies have the form

$$p_t(w \mid x) = \frac{1}{Z_\lambda(x)} p_0(w) e^{-\lambda(t)d(w,x)}, \quad (3)$$

where $Z_\lambda(x)$ is a normalizing constant. We posit that time-varying function $\lambda(t)$ has the form $\lambda(t) = \lambda_0 t$ for constant λ_0 (thus implicitly defining the time-varying depth constraint $C(t)$). This stipulation is inspired by models of perception as noisy evidence accumulation (Bicknell & Levy, 2010; Bogacz et al., 2006; Norris & Kinoshita, 2008), where a prior $p_0(w)$ is continually updated in response to noisy perceptual evidence. The final interpretation policy which we use in this paper is

$$p(w \mid x) \propto p_0(w) e^{-\lambda_0 t d(w,x)}. \quad (4)$$

To give some intuition for model behavior: at time $t = 0$, the interpretation policy reduces to $p_t(w \mid x) = p_0(w)$, meaning that a comprehender's interpretations are based on prior expectations, taking into account no information from the input. As the time $t \rightarrow \infty$, the policy $p_t(w \mid x)$ concentrates all probability mass on interpretations with minimum distortion, that is, on veridical interpretations, at which point the processing depth is equal to the surprisal of the input, $D(t) = -\log p(x)$, indicating that all the information in the input x has been processed.

Application to language comprehension We represent the inputs x and interpretations w as strings. We use a distortion

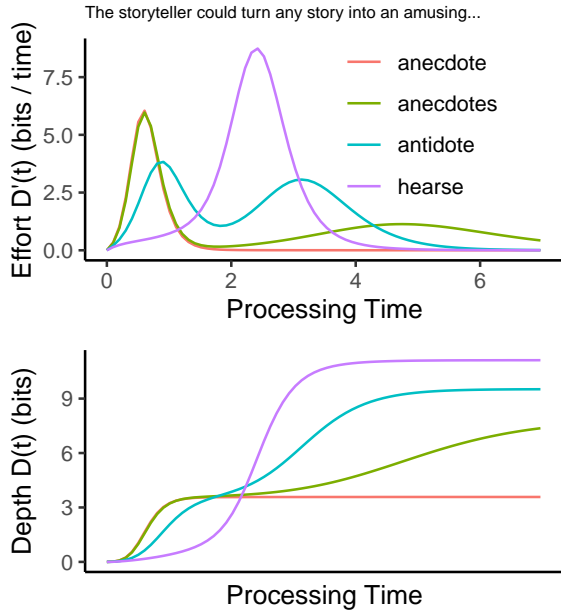The storyteller could turn any story into an amusing...

Figure 3: Processing timecourses for four different inputs in the given context. "Anecdote" is the control and the most likely completion. "Hearse" represents a semantic anomaly, "anecdotes" represents a syntactic anomaly, and "antidotes" represents a recoverable anomaly—semantically anomalous input that can be easily mistaken for a more likely input.

measure that reflects a mixture of semantic and form-based distance between w and x:

$$d(w,x) = d_\varphi(w,x) + \gamma d_\sigma(w,x), \qquad (5)$$

where d_φ is a form-based phonological or orthographic distance metric, d_σ is a semantic distance metric, and γ is a scalar controlling the relative importance of semantic as opposed to form-based distance. We implement the form-based distance as orthographic edit distance and semantic distance as cosine distance between GPT-2 word embeddings (Radford et al., 2019). For the default policy $p_0(w)$, we use the GPT-2 language model (Radford et al., 2019).

Figure 2 shows model behavior using this setup for the last word in Example (1). For this example, we use only form-based distortion (that is, $\gamma = 0$), and the set of possible interpretations w ranges over the 10,000 most frequent words in SUBTLEXus (Brysbaert & New, 2009). Initially, when no perceptual input is taken into account, "story" is the most likely interpretation, as it is the most likely continuation in the context under the GPT-2 language model. With short processing time, "anecdote" appears most likely, as it is highly expected in context and also close in distortion to the true input. With deeper processing, all probability mass is concentrated asymptotically on the veridical "antidote". The shift from "anecdote" to "antidote" happens when processing depth reaches around 5 bits.

Processing effort So far our model describes the evolution of a comprehender's interpretations of their input with time. We require a link between this model and measures of processing effort, such as EEG amplitude and reading time. We propose that processing effort corresponds to *change* in processing depth: it requires effort to increase depth. We define the instantaneous **effort** at time t as the rate of change of processing depth $D(t)$, given by the time derivative $D'(t) = \frac{d}{dt}D(t)$.

Example timecourses for depth $D(t)$ and effort $D'(t)$ are shown in Figure 2, bottom. Here, the early concentration of probability mass on "anecdote" is reflected in an initial pulse in the effort $D'(t)$. The later shift of probability mass from "anecdote" onto the veridical "antidote" creates a second pulse. Thus, processing *appears* to proceed in two distinct stages: first a 'shallow' stage of processing, followed by a later 'deep' error correction. This behavior is in fact generated by a single process of continuously increasing processing depth with time.

Link to EEG measures The idea of shallow-to-deep language processing has also been used to explain EEG signals of language comprehension, with the N400 signal a function of effort in shallow processing and the later P600 signal a signal of anomalies detected through deep processing (Hagoort et al., 1993; Hoeks et al., 2004; Ito et al., 2016; Kim & Osterhout, 2005; Kolk et al., 2003; Kuperberg, 2007; Kuperberg et al., 2020; Kutas & Hillyard, 1980; Van Herten et al., 2005, 2006; Van Petten & Luka, 2012). In support of this idea, anomalous words such as those in Example (1), which are surprising but close in form to a plausible alternative, yield a biphasic N400 and P600 effect (Ryskin et al., 2021).

To link our measures to EEG data, we posit the voltage $V(t)$ of EEG signal at time t is proportional to the effort $D'(t)$ modulated by a carrier wave with angular frequency ω and phase ϕ, as

$$V(t) \propto -D'(t)\sin(\omega t + \phi). \qquad (6)$$

In this view, the N400 and P600 signals are manifestations of a singular underlying process of increasing processing depth. N400 arises when there is a large early pulse in effort, and P600 arises when there is a late pulse.

In support of this idea, Figure 3 shows processing timecourses for four different inputs used in Ryskin et al., 2021: a control input "anecdote", a syntactically anomalous input "anecdotes", a semantically anomalous input "hearse", and a recoverable semantic anomaly "antidote" that is phonologically close to the control. The semantic anomaly yields a single, large, delayed pulse in processing effort. The delay happens because early processing for this input fails to concentrate probability mass on any particular interpretation—"hearse" is initially so unlikely under $p_0(w)$ that a great deal of evidence has to be accumulated for its probability to rise. In contrast, the syntactic and recoverable semantic anomalies create two pulses of processing effort. For the syntactic anomaly, the timecourse of processing is nearly identical to
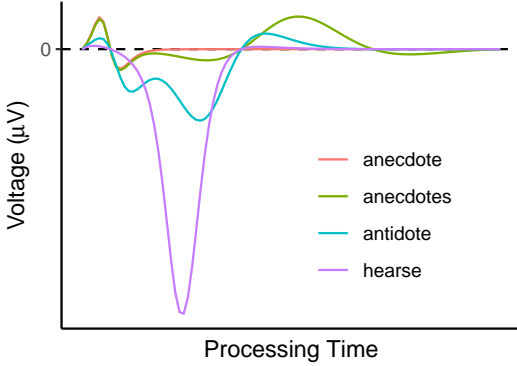
Figure 4: Simulated EEG signal corresponding to the processing timecourses in Figure 3, using Eq. 6 with $\omega = 1$ and $\phi = -2/3$.

the control input up to a depth of around 5 bits, at which point there is a very late and temporally diffuse second pulse of processing effort, corresponding to the P600 signal. Mapping the processing effort onto predicted voltages following Eq. 6 in Figure 4, we predict (1) early negativity for the semantic and recoverable semantic anomalies, and (2) late positivity for the syntactic and recoverable semantic anomalies. This pattern matches the empirical findings.

Link to reading times   To capture reading times, we posit that comprehenders move to the next word when they judge that deeper processing is no longer necessary. Concretely, we model reading times as the minimum amount of time needed for processing effort to fall below a threshold ε. It is thus possible that a reader may move on to the next word before fully resolving the current word, as long as the *change* in processing depth with time is small enough. For example, in the processing timecourse for the syntactically anomalous "anecdotes" in Figure 3, a reader may move on during the trough between the two pulses in processing effort, if the processing effort falls below the threshold ε at this point. In that case the reader would have a lingering misinterpretation.

Relation to noisy-channel models   Our shallow interpretation policy generalizes noisy channel models of language comprehension (Gibson et al., 2013; Poppels & Levy, 2016; Ryskin et al., 2021). The policy reproduces the predictions of noisy-channel models in the special case when $\lambda(t) = 1$ and the distortion metric is equal to a log-likelihood under a noise model, i.e. $d(w, x) = -\ln p_N(x \mid w)$ for some noise model p_N. It differs from noisy channel models in its cognitive interpretation and in its generality: in the shallow processing model, the distortion $d(w, x)$ need not be a log-likelihood (that is, it need correspond to a normalized probability $p_N(x \mid w)$). Increasing t in Eq. 4 corresponds to accumulation of samples of evidence in a noisy-channel model of perception (Bicknell & Levy, 2010; Norris & Kinoshita, 2008).

Study 1: N400, P600, and biphasic EEG signals

Exp	Condition	Context	Target	ERPs
Ito-16	SemRelated	The student is	page	Reduced N400
	FormRelated	going to the li-	hook	Biphasic
	Unrelated	brary to bor-	sofa	N400
	Control	row a . . .	book	NA
Ryskin-21	Semantic	The storyteller	hearse	N400
	Syntactic	could turn any	anecdotes	P600
	SemCrit	incident into	antidote	Biphasic
	Control	an amusing . . .	anecdote	NA

Table 1: List of conditions, sample sentences and ERP patterns in dataset.

Dataset   We validate our theory on two EEG experiments, featuring N400 effect, P600 effect and biphasic effect. Table 1 shows a list of conditions with sample stimuli and empirical ERP patterns across experiments. The ERP effects in the experimental conditions are all calculated in terms of differences to the EEG amplitude in the control condition. The experiment *Ito-16* includes three different kinds of semantic violations (Ito et al., 2016) that are semantically related (*SemRelated*), orthographically related (*FormRelated*) or unrelated (*Unrelated*) to the predicable target. The N400 effects in *FormRelated* and *SemRelated* conditions are smaller than in the *Unrelated* conditions, and only *FormRelated* condition elicits a P600 effect. The experiment *Ryskin-21* introduces semantic violations (*Semantic* condition), syntactic violations (*Syntactic* condition) and semantic violations that are phonologically similar to the predictable control target (*SemCrit* condition). There is a biphasic N400–P600 effect in the *SemCrit* condition, but the effect size is smaller than the N400 in the *Semantic* condition and P600 in the *Syntactic* condition.

Implementation   We calculate interpretation policies over time for the critical word in each sentence, with the set of alternative interpretations w consisting of the words in all conditions. We set the parameter γ by visual inspection, obtaining $\gamma = 10$ for Ryskin-21 and $\gamma = 6$ for Ito-16.

Results   Figure 5 shows that our model simulates the observed EEG patterns in both Ito-16 and Ryskin-21. In Ito-16, our model predicts a graded N400 effects elicited by three kinds of semantic violations (*Unrelated* > *SemRelated* > *FormRelated*), where semantic violations that are either orthographically nor semantically related to the predicted word trigger the largest N400 responses. Our model also predicts a greater P600 amplitude in *FormRelated* than other conditions. In Ryskin-21, our model predicts the order the effect size across conditions: a graded N400 effect (*Semantic* > *SemCrit* > *Syntactic*) and a graded P600 effect (*Syntactic* > *SemCrit* > *Semantic*).

The model behavior can be explained in terms of the different rates of probabilistic distribution shift: the semantically

anomalous conditions with a large N400 effect tend to have the fastest probability update, and anomalous conditions with a large P600 effect (e.g. *Syntactic* condition in Ryskin-21; *FormRelated* condition in Ito-16) has the lowest probability update rate. The different probability shift rate results in the increase of instantaneous effort to appear early or delayed.

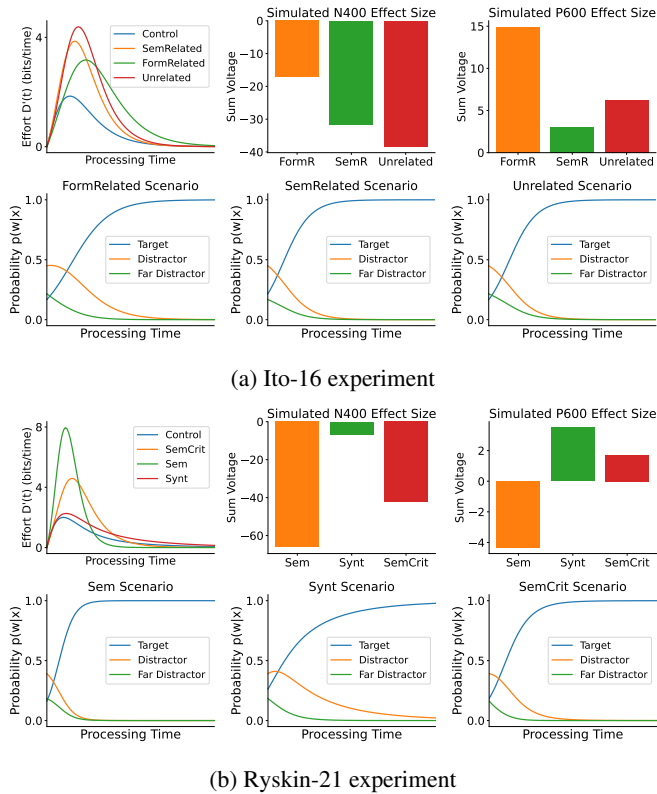(a) Ito-16 experiment

(b) Ryskin-21 experiment

Figure 5: Simulation results for EEG experiments. **Left.** Instantaneous processing effort ($D'(t)$) **Middle.** Simulated N400 effect size. **Right.** Simulated P600 effect size.

Study 2: Garden path reading times

Dataset We used a set of three classic garden path constructions from large-scale syntactic ambiguity processing benchmark, consisting of self-paced reading time data from 2000 participants (Huang et al., 2023). The constructions include: Main verb/ reduced relative clause garden path (MV/RR), Direct object/ sentential complement garden path (NP/S), Transitive/ intransitive garden path (NP/Z) (see Table 2). There are 24 lexically matched item sets for each construction. Each set has an experimental sentence with a tempory syntactic ambiguity and an unambiguous control sentence. These constructions generate reliable garden path effects (Bever, 1970; Frazier, 1979; Grodner et al., 2003; Sturt et al., 1999), but differ in the magnitude of the effect (Huang et al., 2023; Van Schijndel & Linzen, 2021). In particular, NP/Z constructions have the largest effect size at the critical region, followed by MV/RR and NP/S: see Figure 6(a).

Type	Sentence
MV/RR	The little girl fed the lamb **remained** relatively calm. . .
	The little girl who was fed the lamb *remained* relatively calm. . .
NP/S	The little girl found the lamb **remained** relatively calm.
	The little girl found that the lamb *remained* relatively calm. . .
NP/Z	When the little girl attacked the lamb **remained** relatively calm. . .
	When the little girl attacked, the lamb *remained* relatively calm. . .

Table 2: A list of garden path constructions and example sentences. Critical words are marked in **bold**. The critical words in garden path sentences are underlined, and the control counterparts are *italic*.
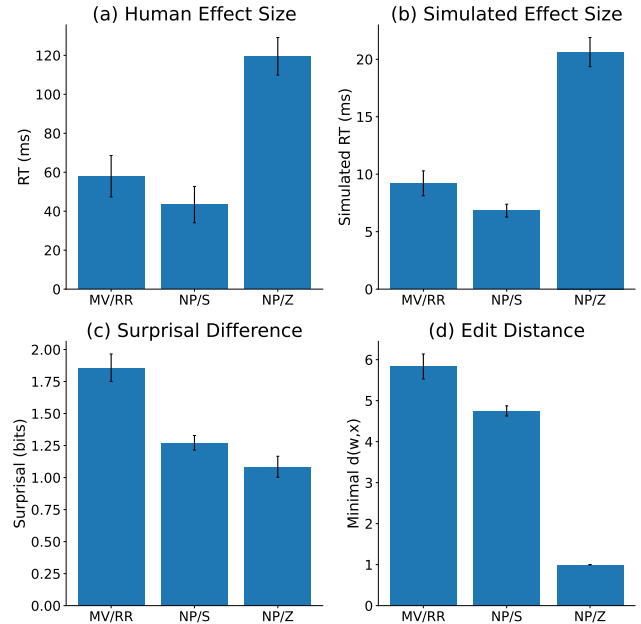

Figure 6: **(a)** Average effect size (RT difference to unambiguous control) at critical region estimated from human experiment. **(b)** Average effect size estimated from model simulation. **(c)** Average surprisal difference between experimental and control sentences. **(d)** Average edit distance between presented sentence and closest alternative. Error bars show standard errors.

Implementation We consider interpretations w ranging over the entire sentence, rather than only the critical word, with a set of alternatives consisting of all sentences in a lexically matched set across different constructions as possible interpretations given a single sentence. This simulates the scenario where a comprehender is re-evaluating and updat-

ing their beliefs about the entire sentence in parallel (Wen et al., 2021a, 2021b). The prior $p(w)$ is calculated as the sum of conditional log probability for every word in the sentence from GPT-2. We set $\gamma = 0$ because all alternatives are lexically matched, and set the reading time threshold to $\varepsilon = .001$.
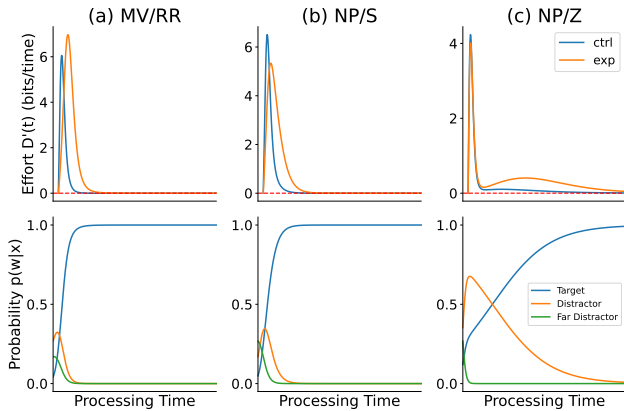

Figure 7: **Top.** Instantaneous processing effort over time for (a) MV/RR; (b) NP/S and (c) NP/Z constructions. The dotted red line indicates the reading time threshold $\varepsilon = .001$. **Bottom.** Probabilities of three different interpretations w given an experimental sentence (*Target*). *Distractor* is the counterpart control sentence and *Far Distractor* is a sentence with a different syntactic construction.

Results Figure 6 shows that our simulation replicates the relative magnitude of the three garden path effects. This result cannot be explained by the surprisal difference between experimental and control conditions, where the MV/RR has the largest surprisal difference followed by NP/S and NP/Z, nor by the distortion (edit distance) alone: see Figure 6(c–d).

The model captures the reading time differences among garden paths because the reading time is affected by the comparative strength of different candidate interpretations, which is jointly decided by both prior expectations and similarity to the input. In Figure 7(a–b), we see that for MV/RR and NP/S, the veridical interpretation initially has a very low probability, but the probability increases rapidly because the distance to the more-likely control interpretation is large. On the other hand, for NP/Z—shown in Figure 7(c)—the veridical interpretation is not extremely low probability initially, but the probability for this interpretation grows very slowly because there is only weak evidence for it: the distance between the veridical input and the closest distractor (the unambiguous variant with a comma) is very small.

Discussion

We have proposed a unified computational model of language comprehension as a function of depth of processing. In the model, shallow interpretations are generated based on a trade-off between accuracy and the amount of information

extracted from perceptual input. We quantify reading time and EEG signals as two different indices of changes in comprehenders' beliefs. Our model provides a simple and transparent explanation for the order of effect size across three different types of garden path constructions, and successfully simulates certain ERP effects.

Our model quantifies the well-known hypothesis of a heuristic "shallow processing" mechanism in a time-sensitive and dynamic way. Rather than positing two discrete stages of shallow and then deep processing, or that processing involves separate streams of heuristic and veridical input (Kim & Osterhout, 2005; Van de Meerendonk et al., 2011; Van Herten et al., 2005), our model posits a single interpretation process with continuously increasing in processing depth. For certain inputs, this process creates dynamics with two pulses of processing effort, corresponding to formation of a shallow interpretation followed by error detection or correction.

Our model enhances theories of processing difficulty based on expectations by introducing a new factor in the form of distortion. Our model's indices of processing effort are closely related to the surprisal of the input, which has been proposed as a predictor of reading times (Hale, 2001; Levy, 2008) and EEG amplitudes (Frank et al., 2015; Michaelov et al., 2024). The veridical interpretation's initial probability at time $t = 0$ is determined entirely by prior expectations, and the total cumulative effort to process a word is equal to its surprisal. Our model differs from Surprisal Theory in that reading time is also influenced by distances to distractor interpretations. These distances determine the *rate* at which the probability of the correct interpretation increases, yielding potentially dramatic slowdowns when that rate is small. Our model is also compatible with theories of ERPs where the N400 reflects the surprisal of a likely but non-veridical interpretation, P600 reflects reconciliation of interpretations, and the total cumulative amplitude reflects surprisal (Li & Ettinger, 2023; Li & Futrell, 2023).

While we have demonstrated initial success for our model, it has some limitations. Most serious is the representation of possible interpretations w as strings, with the concomitant use of edit distance (augmented with semantic distance) for the distortion. In a more complete model implementation, it is possible that interpretations w may range over structured objects such as syntactic treelets, with distortion reflecting a structure-sensitive distance function among these objects. Future work will determine the proper structure and distortion measure for comprehenders' interpretations.

Our model and results provide evidence that processing constraints in language comprehension can be captured information-theoretically in terms of rational use of computational resources (Hahn et al., 2022), joining with similar models in the domains of language production (Futrell, 2023), semantics (Imel & Steinert-Threlkeld, 2022; Steinert-Threlkeld, 2020; Zaslavsky et al., 2018), and pragmatics (Zaslavsky et al., 2020; Zhou et al., 2022), as well as other areas of cognition (Arumugam et al., 2022; Bhui et al., 2021).

References

Arumugam, D., Ho, M. K., Goodman, N. D., & Van Roy, B. (2022). On rate-distortion theory in capacity-limited cognition & reinforcement learning. *arXiv preprint arXiv:2210.16877*.

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & cognition*, *21*(4), 477–487.

Bever, T. G. (1970). The cognitive basis for linguistic structures. *Cognition and the development of language*.

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21.

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, 1168–1178.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*(4), 700.

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.

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified account of numerosity perception. *Nature Human Behaviour*, *4*(12), 1265–1272.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive psychology*, *42*(4), 368–407.

Cover, T. M., & Thomas, J. A. (2006). Elements of information theory second edition solutions to problems. *Internet Access*, 19–20.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive psychology*, *47*(2), 164–203.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, *11*(1), 11–15.

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of psycholinguistic research*, *30*, 3–20.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

Fillenbaum, S. (1971). Processing and recall of compatible and incompatible question and answer pairs. *Language and Speech*, *14*(3), 256–265.

Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology*, *102*(4), 574.

Frank, S., Otten, L., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, *140*, 1–11.

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies.* University of Connecticut.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

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. (2023). Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, *120*(39), e2220593120.

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.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the p600. *Language and cognitive processes*, *25*(2), 149–188.

Grodner, D., Gibson, E., Argaman, V., & Babyonyshev, M. (2003). Against repair-based reanalysis in sentence comprehension. *Journal of Psycholinguistic Research*, *32*, 141–166.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and cognitive processes*, *8*(4), 439–483.

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.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the north american chapter of the association for computational linguistics*.

Hoeks, J. C., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive brain research*, *19*(1), 59–73.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2023). Surprisal does not explain syntactic disambiguation difficulty: Evidence from a large-scale benchmark.

Imel, N., & Steinert-Threlkeld, S. (2022). Modal semantic universals optimize the simplicity/informativeness trade-off. *Semantics and Linguistic Theory*, *1*, 227–248.

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of memory and language*, *52*(2), 205–225.

Kolk, H. H., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and language*, *85*(1), 1–36.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, *1146*, 23–49.

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the n400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35.

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological psychology*, *11*(2), 99–116.

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

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, *6*(2), 279–311.

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. (2023). A decomposition of surprisal tracks the n400 and p600 brain potentials. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

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.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of language*, 1–29.

Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., & Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, *118*(49), e2025993118.

Norris, D., & Kinoshita, S. (2008). Perception as evidence accumulation and bayesian inference: Insights from masked priming. *Journal of Experimental Psychology: General*, *137*(3), 434.

Poppels, T., & Levy, R. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. *CogSci*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

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.

Sanford, A. J. (2002). Context, attention and depth of processing during interpretation. *Mind & language*, *17*(1-2), 188–206.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.

Steinert-Threlkeld, S. (2020). Quantifiers in natural language optimize the simplicity/informativeness trade-off. *Proceedings of the 22nd Amsterdam colloquium*, 513–522.

Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, *40*(1), 136–150.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental psychology: Learning, memory, and Cognition*, *19*(3), 528.

Van de Meerendonk, N., Indefrey, P., Chwilla, D. J., & Kolk, H. H. (2011). Monitoring in language perception: Electrophysiological and hemodynamic responses to spelling violations. *Neuroimage*, *54*(3), 2350–2363.

Van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. In *Reading as a perceptual process* (pp. 621–648). Elsevier.

Van Herten, M., Chwilla, D. J., & Kolk, H. H. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of cognitive neuroscience*, *18*(7), 1181–1197.

Van Herten, M., Kolk, H. H., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive brain research*, *22*(2), 241–255.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190.

Van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988.

Wen, Y., Mirault, J., & Grainger, J. (2021a). Fast syntax in the brain: Electrophysiological evidence from the rapid parallel visual presentation paradigm (rpvp). *Journal of Ex-*

perimental Psychology: Learning, Memory, and Cognition, *47*(1), 99.

Wen, Y., Mirault, J., & Grainger, J. (2021b). The transposed-word effect revisited: The role of syntax in word position coding. *Language, Cognition and Neuroscience*, *36*(5), 668–673.

Zaslavsky, N., Hu, J., & Levy, R. P. (2020). A rate-distortion view of human pragmatic reasoning. *arXiv preprint arXiv:2005.06641*.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Zhou, I., Hu, J., Levy, R., & Zaslavsky, N. (2022). Teasing apart models of pragmatics using optimal reference game design. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).