Generalized Measures of Anticipation and Responsivity in Online Language Processing

Mario Giulianelli   **Andreas Opedal**   **Ryan Cotterell**
{mario.giulianelli,andreas.opedal,ryan.cotterell}@inf.ethz.ch

ETH *zürich*

Abstract

We introduce a generalization of classic information-theoretic measures of predictive uncertainty in online language processing, based on the simulation of expected continuations of incremental linguistic contexts. Our framework provides a formal definition of anticipatory and responsive measures, and it equips experimenters with the tools to define new, more expressive measures beyond standard next-symbol entropy and surprisal. While extracting these standard quantities from language models is convenient, we demonstrate that using Monte Carlo simulation to estimate alternative responsive and anticipatory measures pays off empirically: New special cases of our generalized formula exhibit enhanced predictive power compared to surprisal for human cloze completion probability as well as ELAN, LAN, and N400 amplitudes, and greater complementarity with surprisal in predicting reading times.

⭘ https://github.com/rycolab/
generalized-surprisal

1 Introduction

The prediction of upcoming linguistic units is posited to play a key role in human language comprehension (Federmeier, 2007; Willems et al., 2016; Goldstein et al., 2022). One fruitful method of operationalizing human uncertainty over predictions is through information-theoretic measures. Because human predictive mechanisms leave behavioral and neural traces that are observable during reading and listening (Kutas and Hillyard, 1984; Van Berkum et al., 2005; Forseth et al., 2020), the most common method of vetting an information-theoretic measure of predictive uncertainty is by examining its relationship with such traces. Beyond simply yielding good correlates, information-theoretic measures often provide insight into the human prediction mechanism, and they are thus central to much cognitive and neurobiological research on human language processing (Monsalve et al., 2012; Armeni et al., 2017; Wilcox et al., 2023).

In the domain of sentence processing, there are two commonly deployed information-theoretic measures of predictive uncertainty. Both assume that the comprehender implicitly maintains a probability distribution over upcoming sequences of linguistic units. The first, and most prominent, is the surprisal of a unit given its preceding context (Hale, 2001), while the other is entropy (Hale, 2003, 2006). In broad strokes, surprisal tells us how likely the next unit is in the given context and is a good example of a **responsive** measure, i.e., a measure that quantifies a response to the *next* unit. In contrast to surprisal, entropy is solely a function of the context, as it tells us the uncertainty over the range of possible upcoming linguistic units. Thus, entropy is an example of an **anticipatory** measure, i.e., a measure that anticipates a response to the next unit without knowing its identity. In the specific case of next-symbol entropy (Frank, 2013; Pimentel et al., 2023), it is the expected value of the next-symbol surprisal so it comes with a natural interpretation of the expected response.

Estimates of surprisal and entropy based on neural language models have demonstrated significant predictive capacity for a wide variety of neural and behavioral data collected using self-paced and eye-tracked reading experiments (Goodkind and Bicknell, 2018; Wilcox et al., 2023), as well as EEG (Merkx and Frank, 2021; Michaelov et al., 2024), fMRI (Shain et al., 2020; Bhattasali and Resnik, 2021), and ECoG imaging (Schrimpf et al., 2021), along with explicit grammaticality and acceptability ratings (Lau et al., 2017; Wallbridge et al., 2022). Despite surprisal and entropy's empirical success, there is increasing interest in defining and evaluating alternative measures. Examples include measures designed to disentangle different dimensions—e.g., lexical versus syntactic—of uncertainty (Roark et al., 2009; Arehalli et al., 2022; Giulianelli et al., 2023) or to quantify uncertainty over spans larger than a single unit (Aurnhammer and Frank, 2019; Giulianelli et al., 2024b). While estimating surprisal and entropy from neu-

ral language models is convenient, experimenters should chart new territory by designing their own measures, rethinking and enhancing established information-theoretic quantities to capture overlooked aspects of online language processing.

In this paper, we introduce a new framework, termed generalized surprisal, for responsive and anticipatory models of online language processing. This framework encompasses existing information-theoretic measures as special cases and offers a new method to develop novel ones. We begin by deriving a generalization of surprisal, demonstrating that it corresponds to an expectation over continuations of a linguistic context (§2). We then show how many existing measures can be seen as special cases of our generalized formula, and we propose new special cases, such as sequence-level entropy and next-symbol information value (§3). Because some special cases cannot be calculated in closed form, we must rely on a sampling procedure. This introduces a trade-off between runtime and variance, which we analyze empirically in §5. Finally, we evaluate all special cases as predictors of neural and behavioral data collected in experiments with human participants (§6).

We present several new findings, including: (1) contextual probability predicts human cloze completions better than surprisal, while surprisal is a better predictor of human predictability ratings; (2) information value predicts N400 better than surprisal, which is among the go-to predictors for this ERP component (DeLong et al., 2005; Frank et al., 2015; Michaelov et al., 2024); (3) sequence-level entropy, introduced in this paper, is the sole significant predictor of ELAN; and (4) different responsive measures predict ERP amplitudes at varying time windows after stimulus onset.

2 Generalized Surprisal

This section introduces our framework. First, we establish some key notation and definitions. Then, we present a decomposition of surprisal, which motivates our definition of generalized surprisal.

2.1 Language Modeling

An **alphabet** Σ is a finite, non-empty set of symbols, and its Kleene closure Σ^* is the set of all strings formed by concatenating symbols in Σ, including the empty string ε.[1] The set of all strings

Σ^* is partially ordered by the **prefix relation** $\preceq$, defined as follows: $\boldsymbol{w} \preceq \boldsymbol{w}' \iff \exists \boldsymbol{v} : \boldsymbol{wv} = \boldsymbol{w}'$. As is easy to see, $\preceq$ is reflexive and transitive, but not symmetric. A **language model** p is a distribution over strings Σ^*. A common quantity derived from a language model is the **prefix probability**, defined as

$$\pi_p(\boldsymbol{w} \mid \boldsymbol{c}) \stackrel{\text{def}}{=} \sum_{\boldsymbol{v} \in \Sigma^*} p(\boldsymbol{wv} \mid \boldsymbol{c}). \qquad (1)$$

In words, Eq. (1) tells us the probability of the event that a string sampled from p starts with $\boldsymbol{w}$. Crucially, this is different from the probability $p(\boldsymbol{w} \mid \boldsymbol{c})$ that the string *is* identically $\boldsymbol{w}$.

The Human Language Model. So far, we have used the symbol p to refer to an arbitrary language model. However, in the context of cognitive modeling, we are interested in a hypothetical construct model—the human language model p_H. Because the true human language model is unknown, we must approximate it via another language model p. To the extent that p is close to p_H (under some notion of distance between distributions), we would expect estimates derived from p to be a reliable proxy of the probabilities prescribed by the human language model. In our experiments, we will use a model p parameterized by a transformer neural network, which was shown to closely approximate p_H in a series of psycholinguistic studies (Schrimpf et al., 2021; Oh and Schuler, 2023; Shain et al., 2024, *inter alia*).

2.2 Generalizing Surprisal

The **surprisal** of a **target** $\boldsymbol{w} \in \Sigma^*$ given a **context** $\boldsymbol{c} \in \Sigma^*$ is defined as $\iota_p(\boldsymbol{w}; \boldsymbol{c}) \stackrel{\text{def}}{=} -\log \pi_p(\boldsymbol{w} \mid \boldsymbol{c})$.[2] In constructing our framework, we draw inspiration from the following decomposition of surprisal:

$$\iota_p(\boldsymbol{w}; \boldsymbol{c}) \stackrel{\text{def}}{=} -\log \pi_p(\boldsymbol{w} \mid \boldsymbol{c}) \qquad (2a)$$

$$= -\log \sum_{\boldsymbol{v} \in \Sigma^*} p(\boldsymbol{wv} \mid \boldsymbol{c}) \qquad (2b)$$

$$= -\log \sum_{\boldsymbol{v} \in \Sigma^*} p(\boldsymbol{v} \mid \boldsymbol{c}) \mathbb{1}\{\boldsymbol{w} \preceq \boldsymbol{v}\}, \quad (2c)$$

where $\boldsymbol{v}$ is a **continuation** and $\mathbb{1}\{\boldsymbol{w} \preceq \boldsymbol{v}\}$ is an indicator function that returns 1 when $\boldsymbol{w} \preceq \boldsymbol{v}$ is true and 0 otherwise. When viewed through the lens of

---

[1] We use an unbolded font for symbols, i.e., $w \in \Sigma$, and a bolded font for strings $\boldsymbol{w} \in \Sigma^*$. The concatenation of two strings $\boldsymbol{w}$ and $\boldsymbol{v}$ is written as $\boldsymbol{wv}$. The length of a string is the number of symbols it contains and is denoted as $|\boldsymbol{w}|$.

[2] Note that this is not equal to information content in a strict sense, since that would require $\pi_p(\cdot \mid \boldsymbol{c})$ to be a probability distribution over Σ^*, for all $\boldsymbol{c} \in \Sigma^*$.

such a decomposition, surprisal is a marginalization over possible continuations of the context c: it is the negative log-transformed cumulative probability of all the continuations that begin with w.

By writing the sum as an expectation, we can rewrite surprisal as follows

$$\iota_p(\boldsymbol{w}; \boldsymbol{c}) = -\log\left(\underset{\boldsymbol{v}\sim p(\cdot|\boldsymbol{c})}{\mathbb{E}}\left[\mathbb{1}\{\boldsymbol{w}\preceq\boldsymbol{v}\}\right]\right). \quad (3)$$

Our notion of generalized surprisal abstracts Eq. (3) by introducing a **scoring function** g to generalize the indicator function and a **warping function** f to replace the $-\log$ of standard surprisal.

Definition 1 (Generalized Surprisal). *We define a **generalized surprisal** model as the pair* (f, g) *of a warping function* $\mathrm{f}\colon \mathbb{R} \to \mathbb{R}$ *and a scoring function* $\mathrm{g}\colon \Sigma^* \times \Sigma^* \times \Sigma^* \to \mathbb{R}$. *Under a specific model* (f, g), *the generalized surprisal of a target* $\boldsymbol{w}$ *in a context* $\boldsymbol{c}$ *is*

$$\gamma_p(\boldsymbol{w}; \boldsymbol{c}) \overset{\text{def}}{=} \mathrm{f}\left(\underset{\boldsymbol{v}\sim p(\cdot|\boldsymbol{c})}{\mathbb{E}}\left[\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c})\right]\right). \quad (4)$$

The Scoring Function. We call g the *scoring* function because it evaluates each continuation $\boldsymbol{v} \sim p(\cdot \mid \boldsymbol{c})$ against a target $\boldsymbol{w}$, conditioned on a context $\boldsymbol{c}$, yielding a real-valued score. The score quantifies the accuracy of a prediction (or how close the prediction is to the observed target), where the specific notion of closeness is encoded by the experimenter in their definition of g.

The Warping Function. We call f the *warping* function because it applies a transformation to the expected score and thus controls the shape of the distribution of generalized surprisal values for a given score distribution. It is useful to think of the warping function as characterizing the relationship between prediction accuracy and a certain construct or measurement of interest. For instance, in §6.2.1, we show how the same notion of prediction accuracy captured by surprisal's scoring function, $\mathbb{1}\{\boldsymbol{w}\preceq\boldsymbol{v}\}$, is in a nearly linear relationship with human cloze probabilities yet in a logarithmic relationship with human predictability ratings. Much psycholinguistic research aiming to establish the functional relationship between surprisal and processing difficulty (Smith and Levy, 2013; Brothers and Kuperberg, 2021; Wilcox et al., 2023; Shain et al., 2024, *inter alia*) can be seen as testing different hypotheses about the workings of online language processing by instantiating them through varying warping functions.

2.3 Anticipation and Responsivity

An important distinction between various generalized surprisal models is whether they characterize anticipatory or responsive online processes. These notions have been introduced informally by Pimentel et al. (2023). We give a formal definition of anticipation and responsivity below.

Definition 2 (Anticipation and Responsivity). *We call a generalized surprisal model* (f, g) ***anticipatory*** *if* g *is constant in* $\boldsymbol{w}$, *i.e.,* $\forall \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{w}', \boldsymbol{c} \in \Sigma^*$, *we have* $\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = \mathrm{g}(\boldsymbol{v}, \boldsymbol{w}', \boldsymbol{c})$. *Otherwise, we call* (f, g) ***responsive***.

Def. 2 differentiates anticipation, a state of uncertainty over possible outcomes that is fully determined by the context and the processor's language model, from responsivity, which expresses uncertainty for a specific next outcome.

3 Special Cases of Generalized Surprisal

In this section, we introduce concrete special cases of generalized surprisal (Eq. (4)), which we evaluate as predictors of human behavior and neural activity recorded during online language processing. Some of these have been previously used to predict such psycholinguistic data, while others are new. All special cases are designed by varying the three core components of our framework—anticipation vs. responsivity, scoring function, and warping function—and are meant to exemplify how different hypotheses about online language processing can be instantiated as generalized surprisal models.

3.1 Responsive Measures

We start by introducing three responsive generalized surprisal models. These are models (f, g) where the scoring function $\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c})$ is *not* constant in $\boldsymbol{w}$ (see Def. 2).

Surprisal. The first generalized surprisal model we will consider is the pair (f, g) corresponding to standard surprisal (Eq. (2a) and (3)):

$$\mathrm{f}(x) = -\log(x) \qquad (5a)$$
$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = \mathbb{1}\{\boldsymbol{w}\preceq\boldsymbol{v}\}. \qquad (5b)$$

The scoring function captures a binary notion of prediction accuracy, while the logarithmic warping function is classically considered to instantiate a view of online language processing where cognitive costs reflect the magnitude of incremental mental representation updates, which is related logarithmically to prediction accuracy (Levy, 2008).

Probability. Replacing standard surprisal's logarithmic warping function with the identity function yields the contextual probability of the next unit:

$$\mathrm{f}(x) = x \qquad (6a)$$
$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = \mathbb{1}\{\boldsymbol{w} \preceq \boldsymbol{v}\}. \qquad (6b)$$

The identity warping function can be seen as instantiating a *facilitation* view of online linguistic prediction, where prediction accuracy is linearly related to cognitive cost (Brothers and Kuperberg, 2021).

Information Value. By replacing the indicator function of surprisal and probability with a scoring function $\mathrm{d}_{\boldsymbol{c}} \colon \Sigma^* \times \Sigma^* \to \mathbb{R}_{\geq 0}$ that measures a graded, possibly context-sensitive notion of distance between strings, we obtain information value (Giulianelli et al., 2023):[3]

$$\mathrm{f}(x) = x \qquad (7a)$$
$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = \mathrm{d}_{\boldsymbol{c}}(\boldsymbol{v}, \boldsymbol{w}). \qquad (7b)$$

Whilst the use of a binary scoring function follows naturally from the surprisal model of cognitive cost, it results in a relatively simplistic notion of prediction accuracy which conflates different aspects of predictive accuracy and does not take into account the communicative equivalence of predictions and observations. In the information value model of cost (Giulianelli et al., 2023, 2024b), instead, prediction accuracy is a continuous score that quantifies the representational distance between predictions and observations. For instance, if the predicted continuation is syntactically different but semantically equivalent to the observed next unit, this results in high syntactic and low semantic information value.

3.2 Anticipatory Measures

We now move to anticipatory generalized surprisal models, where the scoring function $\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c})$ is constant in $\boldsymbol{w}$; see Def. 2. We are not aware of any theoretical justifications for using non-linear warping functions for anticipatory measures, so all the special cases presented here will use the identity function $\mathrm{f}(x) = x$.

Expected Next-symbol Surprisal. We begin with a measure that was recently proposed to study the effects of anticipatory processing on reading comprehension (Pimentel et al., 2023). This is the expected surprisal over the language model's next-symbol distribution, which is defined as follows:[4]

$$-\sum_{u \in \Sigma} \pi_p(u \mid \boldsymbol{c}) \log \pi_p(u \mid \boldsymbol{c})$$
$$- p(\varepsilon \mid \boldsymbol{c}) \log p(\varepsilon \mid \boldsymbol{c}). \qquad (8)$$

This measure can be obtained by instantiating the following generalized surprisal model:

$$\mathrm{f}(x) = x \qquad (9a)$$
$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = -\sum_{u \in \Sigma} \mathbb{1}\{u \preceq \boldsymbol{v}\} \log \pi_p(u \mid \boldsymbol{c})$$
$$- \mathbb{1}\{\varepsilon = \boldsymbol{v}\} \log p(\varepsilon \mid \boldsymbol{c}). \qquad (9b)$$

This model lends itself to multiple interpretations; see Pimentel et al. (2023, §3) for some proposals. One prominent view is that cognitive resources may be budgeted in advance—before the identity of the next symbol is known—in proportion to the magnitude of the mental representation update the processor expects to sustain once the symbol is observed. As we have seen in §3.1, said magnitude corresponds to the logarithm of the prediction accuracy.

Expected Next-symbol Probability. An alternative view of anticipatory mechanisms is that they allow for preemptive processing of upcoming units, with cognitive costs proportional to the prediction accuracy *in string space*. As seen in §3.1, this view can be expressed by discarding the logarithm:

$$\mathrm{f}(x) = x \qquad (10a)$$
$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = \sum_{u \in \Sigma} \mathbb{1}\{u \preceq \boldsymbol{v}\} \pi_p(u \mid \boldsymbol{c})$$
$$+ \mathbb{1}\{\varepsilon = \boldsymbol{v}\} p(\varepsilon \mid \boldsymbol{c}). \qquad (10b)$$

Under this model, expected prediction accuracy is linearly related to cognitive cost.

Expected Next-symbol Information Value. A third view is that contextual uncertainty increases processing cost by requiring the retention of a

---

[3]More precisely, this is information value with the *mean* as a summary statistic (cf. Giulianelli et al., 2023, §3.1). The ordered pair $(\Sigma^*, \mathrm{d}_{\boldsymbol{c}})$ forms a semi-metric space, satisfying all properties of a metric space except, possibly, the triangle inequality. The distance function $\mathrm{d}_{\boldsymbol{c}}$ may also be constant in $\boldsymbol{c}$.

[4]Note that this expression involves $\pi_p(\cdot \mid \boldsymbol{c})$ rather than $p(\cdot \mid \boldsymbol{c})$ since it quantifies the uncertainty over the *first* symbols of possible continuations, rather than over full continuations. The $p(\varepsilon \mid \boldsymbol{c})$ term is included to account for the possibility that the string ends after $\boldsymbol{c}$, i.e., that the continuation *is* ε.

larger number of competing continuations in memory. While this can be modeled through expected next-symbol surprisal and probability (Pimentel et al., 2023), these measures characterize different continuations as distinct objects. An alternative hypothesis is that maintaining multiple continuations in memory should be less costly if they are representationally similar. This can be expressed through information value's scoring function:

$$\mathrm{f}(x) = x \tag{11a}$$

$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = \mathop{\mathbb{E}}_{\boldsymbol{v}' \sim p(\cdot|\boldsymbol{c})} \mathrm{d}_{\boldsymbol{c}}(v_1, v_1'), \tag{11b}$$

with a representational notion of distance $\mathrm{d}_{\boldsymbol{c}}$.

Entropy. The three anticipatory models presented so far assume that anticipatory processes operate solely over the next symbol. One way to capture uncertainty over sequences of symbols is through the entropy of the language model, i.e.,

$$-\sum_{\boldsymbol{v} \in \Sigma^*} p(\boldsymbol{v} \mid \boldsymbol{c}) \log p(\boldsymbol{v} \mid \boldsymbol{c}). \tag{12}$$

The corresponding generalized surprisal model is

$$\mathrm{f}(x) = x \tag{13a}$$

$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = -\log p(\boldsymbol{v} \mid \boldsymbol{c}). \tag{13b}$$

This model inherits the interpretation of expected surprisal but characterizes anticipation as a process spanning longer time intervals (Hale, 2003).

Expected Information Value. We also consider a second generalized surprisal model that captures sequence-level contextual uncertainty:

$$\mathrm{f}(x) = x \tag{14a}$$

$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = \mathop{\mathbb{E}}_{\boldsymbol{v}' \sim p(\cdot|\boldsymbol{c})} \mathrm{d}_{\boldsymbol{c}}(\boldsymbol{v}, \boldsymbol{v}'), \tag{14b}$$

again, with a representational notion of distance $\mathrm{d}_{\boldsymbol{c}}$. Unlike entropy, which regards expected next strings as distinct, expected information value (Giulianelli et al., 2023, §G) corrects for potential similarity between strings in representational space.

3.3 Monte Carlo Simulation

For tractable estimation of generalized surprisal, which requires taking an expectation over continuations $\boldsymbol{v} \in \Sigma^*$, we use Monte Carlo simulation:

$$\widehat{\gamma}_p(\boldsymbol{w}; \boldsymbol{c}) = \mathrm{f}\left(\frac{1}{N} \sum_{n=1}^{N} \mathrm{g}(\boldsymbol{v}^{(n)}, \boldsymbol{w}, \boldsymbol{c})\right), \tag{15}$$

where $\boldsymbol{v}^{(n)} \sim p(\cdot \mid \boldsymbol{c})$ are obtained via ancestral sampling. If f is continuous, then Eq. (15) is consistent. Moreover, if $\mathrm{f}(x) = x$, the estimator is additionally unbiased. The variance of Eq. (15) depends on the scoring function g and the sample size N; in §5.1, we study the influence of N on the variance of the estimator for different generalized surprisal models.

Some special cases of generalized surprisal—in particular, surprisal, probability, and their expected next-symbol versions—can be computed in closed form, and, for those, we do not need to rely on Monte Carlo simulation.

4 Experimental Setup

Dataset. We use the Aligned dataset (de Varda et al., 2023), which consists of $M = 1726$ target–context pairs from English novels annotated with several different neural and behavioral measurements. We include most of these in our experiments: cloze completions (probability and entropy), predictability ratings, event-related brain potentials (ELAN, LAN, N400, P600, EPNP, PNP), eye-tracked reading times (first-fixation time, first-pass time, right-bounded time), and self-paced reading times. Details on these measurements are given in App. A.1. Each target–context pair $(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})$, termed a **stimulus**, is associated with a real-valued measurement $\psi(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})$, termed **datum**, which is an aggregation of per-subject measurements for that stimulus.[5] In our experiments, we will compute generalized surprisal $\{\widehat{\gamma}_p(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})\}_{m=1}^{M}$ for all stimuli in the dataset, and evaluate it as a predictor for the corresponding data $\{\psi(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})\}_{m=1}^{M}$. The contexts $\boldsymbol{c}$ are strings ranging from 5 to 14 words and targets $\boldsymbol{w}$ are strings corresponding to a single word.[6]

Language Models. We obtain generalized surprisal estimates from GPT-2 Small (Radford et al., 2019) and GPT-Neo 125M (Black et al., 2021). Prior work has shown that, despite exhibiting higher test perplexity, these two models have better psycholinguistic predictive power than larger ones (Oh and Schuler, 2023; Shain et al., 2024). Furthermore, their smaller size incurs a lower computa-

---

[5]For all types of psycholinguistic data except cloze probability and cloze entropy (which are aggregates by definition), a single datum is the average per-subject measurement.

[6]A word is taken to be a contiguous sequence of characters delimited by a white space. Following de Varda et al. (2023), sentence-initial and sentence-final words, as well as words attached to a comma or clitics are excluded.
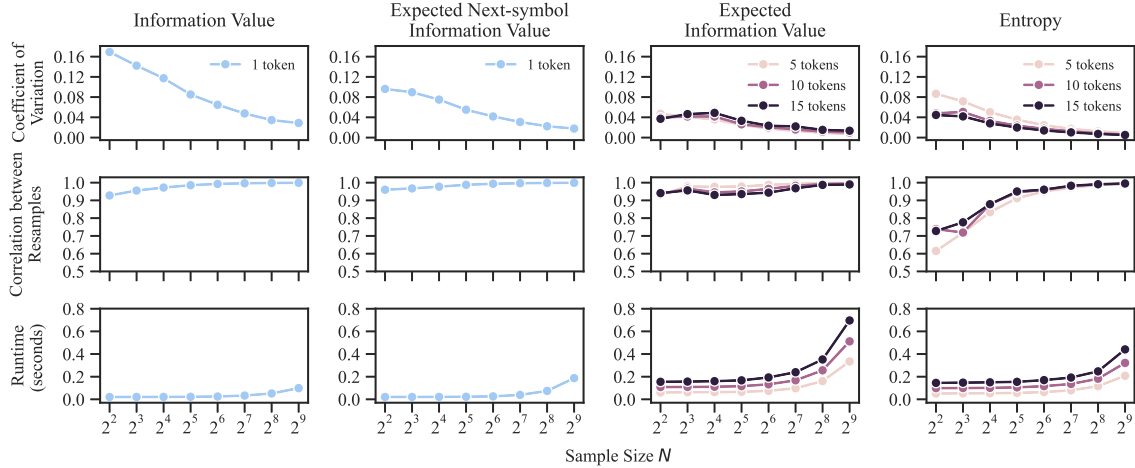
Figure 1: *Coefficient of variation* (top), *correlation between resamples* (center), and *runtimes* (bottom) for sampling-based measures across the stimuli in the Aligned dataset. Confidence intervals (95%) are too narrow to be visible; the horizontal axis is in log scale. The average runtime for the exact metrics (surprisal, probability, expected next-symbol surprisal, and expected next-symbol probability) is 0.002 seconds.

tional cost for sampling. For comparability to prior work (Wilcox et al., 2023; Pimentel et al., 2023; de Varda et al., 2023, *inter alia*), when a word is composed of multiple subword tokens, token-level estimates of generalized surprisal are aggregated (either summed or multiplied, depending on the special case; see App. A.2 for details).

Distance Function. The information value models introduced in §3.1 rely on a function d_c to quantify the distance between two strings v and w in the context of c. Here, we use mean-pooled non-contextual (i.e., layer 0) representation from GPT-2 Small and GPT-Neo 125M as column vectors $r(v), r(w) \in \mathbb{R}^D$, where D is the dimensionality of the representations, and calculate the cosine distance $d_c(v, w) = 1 - \frac{r(v)^\mathsf{T} r(w)}{\|r(v)\|_2 \|r(w)\|_2}$. There exist several other choices for distance and representation functions; see Giulianelli et al. (2023, 2024b) and Meister et al. (2024) for some examples.

5 Empirical Analysis of Measures

We begin our experiments with an empirical analysis of Monte Carlo estimation for the measures presented in §3. As alluded to earlier, sampling-based measures introduce variance that does not arise for measures like surprisal, which can be computed in closed form. This variance can be reduced by increasing the sample size (§5.1), but that incurs additional costs in terms of runtime (§5.2). In this section, we provide an empirical analysis of these properties to gain an understanding of their trade-off. We also investigate the correlations between different measures, in App. B, to assess the extent

to which various theoretical models of anticipatory and responsive processing lead to divergent empirical uncertainty measurements. The analysis in this section is based on GPT-2 Small; results for GPT-Neo 125M are provided in Appendix App. C.

5.1 Variation in Estimates

We use bootstrapping (Efron, 1992) to measure the variance of different estimators. For each stimulus $(w^{(m)}, c^{(m)})$ in the Aligned dataset (§4), given an original sample of size $N \in \{2^j \mid j = 2, 3, \ldots, 9\}$, we obtain $B = 1000$ resamples of the same size by sampling *with replacement*. For entropy and expected information value, to yield tractable estimation, we limit the maximum length in tokens $L \in \{5, 10, 15\}$ of sampled continuations.

In a first analysis, we compute μ_m and σ_m as the mean and standard deviation of a measure of choice across the B resamples and calculate the coefficient of variation $CV_m = \sigma_m / \mu_m$. Fig. 1 (top) shows how average CV, as expected, decreases with N. The maximum sample length L has a limited effect on the CV of sequence-level measures. In a second analysis, to gauge the robustness of a given measure across a dataset of stimuli, we also calculate correlations between different resamples. For each measure, we obtain a matrix of $M \times B$ estimates, and then compute a vector of $\binom{B}{2}$ Pearson correlations between all two-column combinations. Fig. 1 (center) shows the average correlation coefficients as a function of increasing sample sizes. With the exception of entropy, which only reaches near-perfect correlation with $N = 2^7$, all measures show near-perfect correlation already with $N = 2^5$,

indicating that all bootstrapped resamples yield virtually equivalent estimates for the dataset.

In sum, for all measures, both stimulus- and corpus-level variance are contained, they decrease with larger sample sizes, and a sample size as small as 2^5 yields consistent estimates.

5.2 Runtime

We study runtime using the same sample sizes N and maximum sequence lengths L as above, and for the same measures. Fig. 1 (bottom) displays the results. As expected, the runtime increases monotonically in both N and L. Comparing these results to those of the variance analysis suggests a good runtime-variance tradeoff can be obtained with a sample size N between 2^5 and 2^7.

6 Psycholinguistic Predictive Power

We now evaluate the generalized surprisal models introduced in §3 in terms of their predictive power for the neural and behavioral data presented in §4.

6.1 Evaluation

To quantify a measure's predictive power for a given data type $\{\psi(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})\}_{m=1}^{M}$ in the Aligned dataset, we use regression analysis. We compare a regressor that includes baseline predictors for $\{(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})\}_{m=1}^{M}$, the **baseline regressor**, to one that further includes the measure of interest $\{\widehat{\gamma}_p(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})\}_{m=1}^{M}$, the **target regressor**. Reading time regressors include target and baseline predictors not just for the target string but also the previous two words to account for spillover effects (Just et al., 1982; Frank et al., 2013a).

For the experiments on responsive measures (§6.2.1), all regressors include three baseline predictors: the length of the target string $\boldsymbol{w}^{(m)}$, its frequency, and the length of the context string $\boldsymbol{c}^{(m)}$.[7] We call this the **default baseline**. For the experiments on anticipatory measures (§6.2.2), we use an additional baseline. Because expected next-symbol surprisal has proven to be most effective as a predictor when used in conjunction with surprisal (Pimentel et al., 2023), we compare the other, as yet untested, anticipatory measures against a baseline that includes both surprisal and expected next-symbol surprisal, along with the three baseline predictors mentioned above (the **combined baseline**). In the target regressor,

expected next-symbol surprisal is then replaced with another anticipatory target predictor, allowing us to assess the boost, or decrease, in predictive power that results from this substitution.[8]

In our analysis, we only use linear regression so that the functional relationship between the scoring function and the psycholinguistic variable can be unambiguously expressed via the warping function.[9] The regressors are fit with the ordinary least squares method and evaluated on a held-out test set. We quantify the **predictive power** of a measure as the difference in the coefficient of determination R^2 of the target regressor and the baseline regressor, denoted as Δ_{R^2}. The statistical significance of a measure's Δ_{R^2} is assessed via 10-fold cross-validation and permutation tests. A full description of our analysis procedure is given in App. D.1.

6.2 Results

We now present our main results, obtained with GPT-2 Small, $N = 2^9$, and $L = 5$ tokens.

6.2.1 Responsive Measures

The main results for responsive measures are visualized in Figs. 2 and 3. See Figs. 10 to 13 in App. D.2 for further results.

Cloze Completions and Predictability Ratings. These are the two types of psycholinguistic data in the Aligned dataset that more explicitly quantify uncertainty for the upcoming unit. Indeed, predictive power here is 1-2 orders of magnitude larger than for ERPs and reading times (Fig. 2 vs. 12). We find surprisal has the highest Δ_{R^2} for predictability ratings (0.30 ± 0.06), while probability has stronger predictive power for human cloze completion probability (0.48 ± 0.07). This result demonstrates the role of the warping function f in Eq. (4) in fitting a given psycholinguistic construct or data type, even when the scoring function remains unchanged (see Eq. (5b) and (6b)). As further illustrated in Fig. 3, the same binary scoring function provides a good linear fit to human cloze probabilities but not to predictability ratings. In contrast, logarithmic warping of the binary scores, which handles highly surprising outcomes more robustly, results in a better fit to predictability ratings.

---

[7]The length of the target string is measured in characters, the length of the context in words. Frequencies are extracted from the SUBTLEXus (Brysbaert et al., 2012).

[8]App. D.2 also reports results for the default baseline.

[9]Others (e.g., Smith and Levy, 2008, 2013; Goodkind and Bicknell, 2018; Brothers and Kuperberg, 2021; Wilcox et al., 2023) have tried to learn the form of this relationship from data, using generalized additive models (Wood, 2004, 2017).
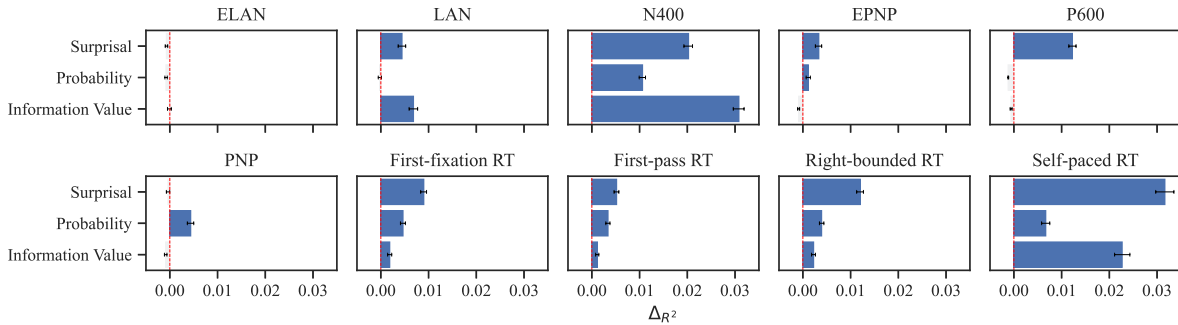
Figure 2: Predictive power (Δ_{R^2}) of responsive generalized surprisal models for event-related potentials and reading times. 95% confidence intervals. Significance color-coded: blue for $p < 0.0001$, gray for $p > 0.01$.
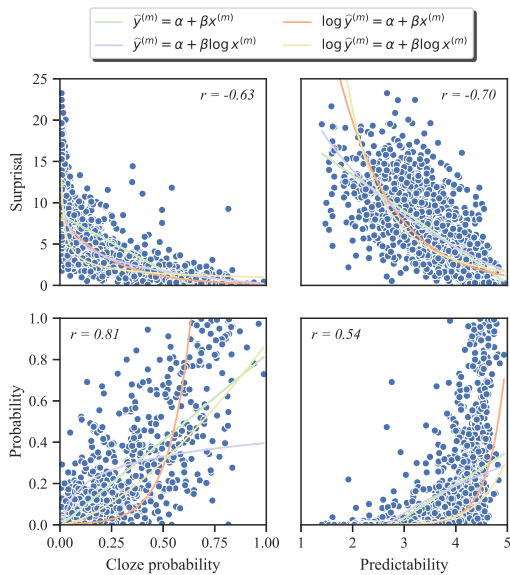

Figure 3: Probability and surprisal against human cloze probabilities and predictability ratings, with Pearson correlation coefficients r and regression lines. For regressions, $x^{(m)} = \gamma(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})$ is the predictor and $y^{(m)} = \psi(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})$ the predicted variable.

Event-related Brain Potentials. Different responsive models—both in terms of warping and scoring function—align with different ERP components (Fig. 2). Surprisal shows higher predictive power for EPNP and P600, probability is the best predictor of PNP amplitudes, and information value of LAN and N400. Information value's high predictive power for N400, a component believed to relate to semantic predictability (Brothers et al., 2020), may be explained in terms of this measure's semantically aware scoring function. Furthermore, the predictive power of different responsive measures appears to correspond to groupings defined by the time windows in which the ERP amplitudes are recorded: LAN and N400 are detected roughly between 300 and 500 ms after the stimulus onset, EPNP and P600 occur between 400 and 700 ms, and PNP is the latest component, recorded be-

tween 600 and 700 ms after the onset (Frank et al., 2015).[10] Overall, this result highlights the importance of having a family of measures at disposal for targeted modeling of various psycholinguistic data.

Reading Times. For both eye-tracked and self-paced reading time data, surprisal demonstrates superior performance among responsive measures. As shown in Fig. 2, probability is the second best predictor for eye-tracked reading times, and information value ranks second for self-paced reading times. These results imply that, among the competing theories we examined, surprisal theory—which posits that cognitive cost is linked to the magnitude of incremental updates in mental representation—provides a better explanation for the traditional notion of cost captured by reading times.

6.2.2 Anticipatory Measures

The main results for anticipatory measures are shown in Fig. 4. See also Figs. 14 to 18 in App. D.2.

Cloze Completions. The most direct quantification of contextual uncertainty in the Aligned dataset is the entropy derived from human cloze completions. Indeed, all anticipatory measures show the strongest predictive power when used in isolation for this data type; see Fig. 14 in App. D.2 (default baseline). Expected next-symbol surprisal and probability have equivalent predictive power for close entropy, followed by expected next-symbol information value. The sequence-level anticipatory measures exhibit lower predictive capacity, with entropy obtaining the lowest Δ_{R^2}. This is not surprising, considering that cloze completions are composed of individual words.

Event-related Brain Potentials. Echoing our findings with responsive measures, expected next-

---

[10]In fact, a fourth cluster consists of ELAN, the earliest ERP component (125–175ms), for which only an anticipatory measure, entropy, shows significant predictive power (Fig. 4).
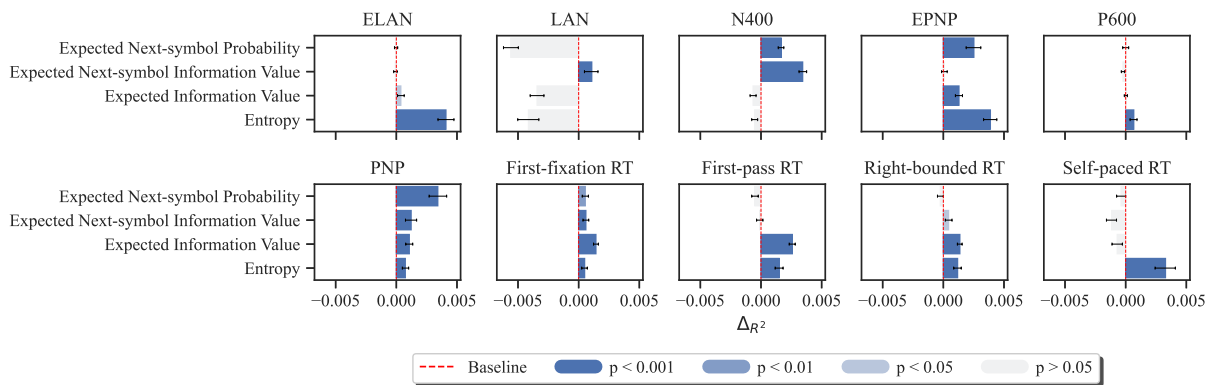
Figure 4: Difference between the predictive power of anticipatory measures used in combination with surprisal vs. expected next-symbol surprisal used in combination with surprisal; 95% confidence intervals. The red dotted line represents the combined baseline regressor. Significance is color-coded, as described in the legend.

symbol information value is the strongest predictor for LAN, and even more so for N400, which emphasizes the connection between N400 and information value's semantically informed scoring function. Entropy also deserves a special mention: It stands out as the sole predictive measure for ELAN, where no other measure—responsive or anticipatory—is significantly predictive. It is also the best predictor of EPNP, with predictive power equivalent to surprisal. A common feature of ELAN and EPNP amplitudes is that they are detected by EEG sensors on frontal scalp regions.

Reading Times. As shown in Fig. 17, where we use the default baseline, reading times are generally less strongly predicted by anticipatory measures in isolation than by responsive ones. However, when considering the combined baseline with surprisal and expected next-symbol surprisal (Fig. 4), we find replacing expected next-symbol surprisal with another anticipatory measure increases predictive power across all data types. The improvements likely stem from the higher complementarity of these measures with surprisal (see App. B).

6.3 Main Findings

Our experiments, conducted across a comprehensive range of psycholinguistic data, demonstrate that different generalized surprisal models—both responsive and anticipatory—provide complementary fit across different data types. For behavioral responses collected in the cloze task (Taylor, 1953), next-symbol probability accurately captures the distribution of human productions, while human predictability ratings are better explained by next-symbol surprisal. For ERP components, the choice of measures has an impact on predictive power

for amplitudes recorded at different onsets. Information value is a stronger predictor of early-onset components, while probability and surprisal are more predictive for late-onset ones. Next-symbol information value, both in its responsive and anticipatory form, is consistently the best predictor for N400, an ERP compoment often predicted using surprisal (Frank et al., 2013b; Michaelov et al., 2024) but also known to be associated with semantic uncertainty (Brothers et al., 2020; Lindborg et al., 2023). On the other hand, sequence-level entropy, a measure whose computation involves long-horizon simulations, is predictive of ERP components in the frontal regions of the scalp, which are thought to be implicated in cognitive or executive control (Alexander et al., 1989; Kandel et al., 2000; Fedorenko et al., 2013). Finally, reading times, both self-paced and eye-tracked, are best predicted by responsive measures, with surprisal emerging as the overall best predictor. However, when comparing models that include surprisal alongside an anticipatory predictor, replacing Pimentel et al.'s (2023)'s expected next-symbol surprisal with one of our alternative anticipatory measures yields significant increases in predictive power across all studied reading time variables.

7 Conclusion

We introduced a generalization over classic information-theoretic measures of predictive uncertainty in online language processing. Our generalized surprisal framework subsumes both responsive and anticipatory measures, including established special cases, but providing a vocabulary and the formal tools for experimenters to design new measures and explain psycholinguistic data of interest.

Limitations

There are several special cases of generalized surprisal that we did not include in our experiments to maintain focus, ensure the interpretability of our results, and keep the scope appropriate for a conference paper. In App. E, we provide a few examples (Rabovsky et al., 2018; Li and Ettinger, 2023; Opedal et al., 2024; Meister et al., 2024).

Other limitations of our study concern the psycholinguistic data under analysis. We consider only English data and native English speakers, and thus, can only draw conclusions about incremental processing of English as L1. Multilingual datasets exist (e.g., Siegelman et al., 2022; Berzak et al., 2022) and should be used in future work to test our findings for other languages as well as speakers of English with a different L1. Furthermore, the linguistic contexts in the analyzed dataset consist of a single sentence. More experimentation is needed to assess the predictive power of our different measures with more complex linguistic contexts such as whole paragraphs and texts, e.g., with the Natural Stories corpus (Futrell et al., 2018), or sequences of conversational turns, which are known to modulate predictive uncertainty in non-trivial ways (Giulianelli and Fernández, 2021; Giulianelli et al., 2021; Tsipidi et al., 2024). Generally speaking, contexts are representations of the current state of the world and can include extra-linguistic information (Ankener et al., 2018; Giulianelli, 2022). Future work should also study responsive and anticipatory linguistic processing modulated by visual cues. For visuo-linguistic contexts, estimates of our generalized formula can be calculated using image-conditioned or video-conditioned LMs.

Finally, while we experiment with increasing the sample size in §5.1, there could be other, more efficient ways to reduce variance. Future work may tackle variance reduction through, for example, importance sampling from altered (e.g., temperature-annealed) language model distributions.

Acknowledgments

References

Michael P. Alexander, D.Frank Benson, and Donald T. Stuss. 1989. Frontal lobes and language. *Brain and Language*, 37(4):656–691.

Christine S. Ankener, Mirjana Sekicki, and Maria Staudte. 2018. The influence of visual uncertainty on word surprisal and processing effort. *Frontiers in Psychology*, 9.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kristijan Armeni, Roel M. Willems, and Stefan L. Frank. 2017. Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, 83:579–588.

Christoph Aurnhammer and Stefan L. Frank. 2019. Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. CELER: A 365-participant corpus of eye movements in L1 and L2 English reading. *Open Mind*, pages 1–10.

Shohini Bhattasali and Philip Resnik. 2021. Using surprisal and fMRI to map the neural bases of broad and local contextual prediction during natural language comprehension. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3786–3798, Online. Association for Computational Linguistics.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow.

Trevor Brothers and Gina R. Kuperberg. 2021. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Trevor Brothers, Eddie W Wlotko, Lena Warnke, and Gina R Kuperberg. 2020. Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, 1(1):135–160.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, 44:991–997.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2023. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*.

Katherine A. DeLong, Laura Quante, and Marta Kutas. 2014. Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, 61:150–162.

Katherine A. DeLong, Thomas P. Urbach, and Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8):1117—1121.

Emanuel Donchin. 1979. Event-related brain potentials: A tool in the study of human information processing. In *Evoked Brain Potentials and Behavior*, pages 13–88. Springer.

Bradley Efron. 1992. Bootstrap methods: Another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

Kara D. Federmeier. 2007. Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4):491–505.

Evelina Fedorenko, John Duncan, and Nancy Kanwisher. 2013. Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 110(41):16616–16621.

Kiefer J. Forseth, Gregory Hickok, Patrick S. Rollo, and Nitin Tandon. 2020. Language prediction mechanisms in human auditory cortex. *Nature Communications*, 11(1):5240.

Stefan L. Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5(3):475–494.

Stefan L. Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013a. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45:1182–1190.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2013b. Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 878–883, Sofia, Bulgaria. Association for Computational Linguistics.

Stefan L. Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Angela D. Friederici and Jürgen Weissenborn. 2007. Mapping sentence form onto meaning: The syntax–semantic interface. *Brain Research*, 1146:50–58.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association.

Mario Giulianelli. 2022. Towards pragmatic production strategies for natural language generation tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7978–7984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mario Giulianelli and Raquel Fernández. 2021. Analysing human strategies of information transmission as a function of discourse context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024a. On the proper treatment of tokenization in psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mario Giulianelli, Sarenne Wallbridge, Ryan Cotterell, and Raquel Fernández. 2024b. Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction. *Preprint*, PsyArXiv:10.31234.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings*

of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

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*.

John Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive science*, 30(4):643–672.

Marcel A. Just, Patricia A. Carpenter, and Jacqueline D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2):228.

Edith Kaan, Anthony R. Harris, Edward Gibson, and Phillip J. Holcomb. 2000. The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, 15:159 – 201.

Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell. 2000. *Principles of Neural Science*, volume 4. McGraw-Hill New York.

Marta Kutas and Steven A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

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

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. 2023. A decomposition of surprisal tracks the N400 and P600 brain potentials. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45.

Alma Lindborg, Lea Musiolek, Dirk Ostwald, and Milena Rabovsky. 2023. Semantic surprise predicts the N400 brain potential. *Neuroimage: Reports*, 3(1):100161.

Clara Meister, Mario Giulianelli, and Tiago Pimentel. 2024. Towards a similarity-adjusted surprisal theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2024. Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, 5(1):107–135.

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408.

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.

Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Wilcox. 2024. On the role of context in reading time prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

Milena Rabovsky, Steven S Hansen, and James L McClelland. 2018. Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9):693–705.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020.

11659

fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307.

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.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863.

Nathaniel J. Smith and Roger Levy. 2008. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30, pages 595–600.

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

Wilson L Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Dianne E Thornhill and Cyma Van Petten. 2012. Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, 83(3):382–392.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Gotlieb Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform information density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Jos JA Van Berkum, Colin M Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort. 2005. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443.

Sarenne Wallbridge, Peter Bell, and Catherine Lai. 2022. Investigating perception of spoken dialogue acceptability through surprisal. In *Interspeech 2022: The 23rd Annual Conference of the International Speech Communication Association*, pages 4506–4510. International Speech Communication Association.

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

Roel M. Willems, Stefan L. Frank, Annabel D. Nijhof, Peter Hagoort, and Antal Van den Bosch. 2016. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Simon N Wood. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

Simon N Wood. 2017. *Generalized additive models: An introduction with R*, 2nd edition. Chapman and Hall/CRC.

A Details of the Experimental Setup

A.1 Measurements in the Aligned Dataset

In our experiments, we use the Aligned dataset (de Varda et al., 2023), which consists of $M = 1726$ target–context pairs from English novels annotated with several different neural and behavioral measurements. We present all types of measurements below.

Cloze Completions. In this incremental version of the cloze task (Taylor, 1953), participants are shown a sentence fragment c where the upcoming target word w is masked, and they are asked to guess what that target word will be. Two data types are derived from cloze completions: the **cloze probability** of the upcoming word, estimated as the Laplace-smoothed (with pseudocount $\alpha = 1$) proportion of participants who pick that word, and the entropy of the Laplace-smoothed cloze distribution (**cloze entropy**).

Predictability Ratings. Participants are presented with a sentence fragment c as well as its corresponding target w. Then, they are asked to rate how likely they think the target is on a Likert scale from 1 to 5 (DeLong et al., 2014). Predictability ratings facilitate the analysis of low-probability words unlikely to appear among cloze completions.

Event-related Brain Potentials (ERPs). ERPs are small voltages generated by participants' neural activity and recorded via an electroencephalogram (Donchin, 1979). Participants are tasked with reading a sentence, then their ERPs are post-processed to obtain word-level measurements, where each word in turn is the target w and its preceding words form the context c. For further details, see Frank et al. (2015). The ERP components analyzed here are the **N400**, often associated with a word's semantic predictability (Brothers et al., 2020), **P600**, implicated in syntactic integration processes (Kaan et al., 2000), (Early) Left Anterior Negativity (**ELAN** and **LAN**), linked to syntactic expectations and working memory (Friederici and Weissenborn, 2007), and (Early) Post-N400 Positivity (**EPNP** and **PNP**), thought to reflect lexical expectations (Thornhill and Van Petten, 2012).

Eye-tracked Reading Times. Participants read a sentence and the time spent looking at each word is recorded, so that each word, in turn, is the target w and its preceding words form the context c. The Aligned dataset contains four types of eye-tracked reading time indices: **first-fixation time**, **first-pass time**, **right-bounded time**, and **go-past** time. For more details, see Frank et al. (2013a). We exclude go-past time, a measurement that includes regressions to previous words, and was found to be noisy in this dataset (de Varda et al., 2023).

Self-paced Reading Times. Participants read a sentence one word at a time in a stationary-window paradigm (Just et al., 1982). The time elapsed between the presentation of a word and the participant's key press to proceed to the next word is the **self-paced reading time**. Contexts c here are taken to be the words preceding the current word w, although these are not physically present on the participants' screen.

A.2 Aggregating Multi-Token Estimates

When a word is composed of multiple subword tokens, token-level estimates are aggregated. For probability-based measures, we naturally multiply token-level estimates following the chain rule. For surprisal and information value, we sum token-level estimates (which, for surprisal, is equivalent to multiplying token probabilities). See Giulianelli et al. (2024a) for a discussion on the proper treatment of tokenization in computational psycholinguistics.

B Correlation Between Measures

To understand the potential overlap or complementarity between measures, we calculate their correlation. App. B.1 and App. B.2 below present results in terms of Pearson correlation and Spearman rank-correlation, respectively.

B.1 Pearson Correlation

Fig. 5 shows Pearson correlation coefficients between estimates obtained with responsive and anticipatory measures. Estimates are computed on the 1726 prefix-continuation pairs of the Aligned dataset (§4) using GPT-2 Small, $N = 2^9$ samples and a maximum sample length of $L = 5$ tokens. Results for GPT-Neo 125M, which follow the same trends, are shown in Fig. 7.

First, we compare MC estimates of surprisal and probability to values computed in closed form, putting our theoretical argument in §2.2 to the test. Probability and MC probability have an almost perfect Pearson correlation ($r = 0.97$). The correlation between surprisal and MC surprisal is also strong, although with a lower coefficient ($r = 0.91$).[11] The rank-correlation between the two pairs of measures is the same, $\rho = 0.96$, as shown in Figs. 6 and 7. This result empirically confirms the formal argument put forward in §2.2: surprisal and probability can be expressed as expectations over continuations of partial linguistic stimuli, scored with an indicator function.

Next, we investigate the relationship between measures of the same type, anticipatory or responsive. This allows us to evaluate whether different models of anticipatory and responsive processing lead to similar, diverging, or complementary measurements. We find that information value correlates more strongly with surprisal than with probability and, in line with this observation, that the highest correlation for anticipatory measures is between expected next-symbol surprisal and expected next-symbol information value. Expected next-symbol probability and expected next-symbol surprisal also correlate strongly ($r = 0.80$). On the other hand, sequence-level anticipatory measures exhibit lower correlations overall; entropy, in particular, correlates only moderately with other measures.

Finally, the anticipatory measure that correlates most strongly with surprisal is expected next-symbol surprisal (see both Figs. 5 and 7 for GPT-2 Small and Figs. 6 and 8 for GPT-Neo 125M, where this is even more evident). This result contributes to explaining our findings for reading times in §6.2.2, where we show that replacing expected next-symbol surprisal with another anticipatory measure, in a model that includes surprisal along with the default baselines, yields an increase in predictive power across all reading time variables.

In summary, this correlation analysis provides empirical support to the theoretical foundations of our generalized framework and confirms that its different special cases quantify alternative notions of responsive and anticipatory processing.

B.2 Spearman Rank-Correlation

As a complement to App. B.1, we show the Spearman rank-correlation coefficients between responsive and anticipatory measures in Figs. 7 and 8. Note the almost perfect correlation between surprisal and MC surprisal, identical to that of probability and MC probability.

C   Variation and Runtime Analysis

In §5 of the main paper, we presented an empirical analysis of Monte Carlo estimation, with the goal of understanding the trade-off between the variance and runtime of each measure's estimator. Here, in Fig. 9, we display the result of this analysis conducted using GPT-Neo 125M.

D   Psycholinguistic Predictive Power

D.1 Statistical Analysis

To evaluate the predictive power of a generalized surprisal model, we use the following procedure. First, we run 10-fold cross-validation, iteratively partitioning the Aligned dataset into a 90% training set and a 10% test set and measuring the coefficients of determination R^2 of the baseline and target regressors on the test set. We repeat this procedure using 100 random seeds, and collect the Δ_{R^2} scores associated with the target predictor. These are the scores that determine the width of the bars and the confidence intervals in Figs. 2, 4, 10 and 12 to 18. Then, to assess the significance of a measure's predictive power—i.e., of

---

[11]We add a small constant ($1e{-}4$) to the expected score in Eq. (3) before taking the logarithm to avoid numerical errors. The Pearson correlation coefficient is mildly sensitive to the choice of constant.

a measure's positive Δ_{R^2} scores—we run paired permutation tests[12] under the null hypothesis that the target regressor's R^2 is smaller or equal to the baseline regressor's R^2 and the alternative hypothesis that the target regressor's R^2 is greater than the baseline regressor's R^2. We use 10,000 resamples and the difference between the sample means as a test statistic. The p-value output by the permutation test (as color-coded or indicated in the captions of Figs. 2, 4, 10 and 12 to 18) is the proportion of the randomized null distribution that is as extreme as the observed value of the test statistic.

For comparison between pairs of regressors which both include target predictors, we use the same procedure, only considering the baseline regressor with each of the target regressors in turn. Our full analysis is implemented in Python and available at `https://github.com/rycolab/generalized-surprisal`.

D.2 Further Results

Responsive Measures.   To complement Figs. 2 and 3 in the main paper, Figs. 10 and 11 show the results for responsive measures obtained with GPT-Neo 125M. Figs. 12 and 13 show the Δ_{R^2} scores of our three responsive measures for cloze probability and predictability.

Anticipatory Measures.   To complement Fig. 4 in the main paper, Fig. 16 shows the results for anticipatory measures used in combination with surprisal, with GPT-Neo 125M estimates. Figs. 14 and 15 show the predictive power of anticipatory measures for cloze entropy.

In §6.2.2 of the main paper (Fig. 4), we evaluate anticipatory measures against a combined baseline that includes surprisal and expected next-symbol surprisal next to the default baseline variables (target length, target frequency, and prefix length). Here, in Figs. 17 and 18, we show the predictive power results for the anticipatory measures against the default baseline.

E   Other Special Cases of Generalized Surprisal

As noted in the Limitations section, there are several special cases of generalized surprisal that we did not include in our experiments to maintain a focused scope for the paper. Below, we provide a few examples.

Semantic Update.   This model, proposed by Rabovsky et al. (2018), is based on changes in neural representations and is given by:

$$\mathrm{f}(x) = x \tag{16a}$$

$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = \mathbb{1}\{w_1 \preceq \boldsymbol{v}\} \sum_{i \in \mathcal{I}} |a_i(w_1) - a_i(c_{|\boldsymbol{c}|})|, \tag{16b}$$

in which $\mathcal{I}$ is an index set corresponding to the neurons at some particular layer in a neural network implementation of a language model, and $a_i(u)$ represents the sigmoid activation of neuron i for symbol u.

Pointwise Mutual Information.   The pointwise mutual information between a word and its context, which under certain conditions yields expressive power equivalent to surprisal (Opedal et al., 2024), can be written as:

$$\mathrm{f}(x) = \log(x) \tag{17a}$$

$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = p(\boldsymbol{c}) \cdot \mathbb{1}\{w_1 \preceq \boldsymbol{v}\}. \tag{17b}$$

Similarity-adjusted Surprisal.   The similarity-adjusted notion of surprisal proposed by Meister et al. (2024) is analogous to information value but uses a similarity function $z_{\boldsymbol{c}} \colon \Sigma^* \times \Sigma^* \to [0, 1]$ as a scoring function and the negative logarithm as a warping function, as given by the following model:

$$\mathrm{f}(x) = -\log(x) \tag{18a}$$

$$\mathrm{g}(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{c}) = z_{\boldsymbol{c}}(\boldsymbol{v}, \boldsymbol{w}) \tag{18b}$$

---

[12]We use the implementation provided by the SciPy library under `scipy.stats.permutation_test`.

Lastly, we note that the decomposition introduced by Li and Futrell (2023) is mathematically equivalent to surprisal and is therefore also captured by our framework.
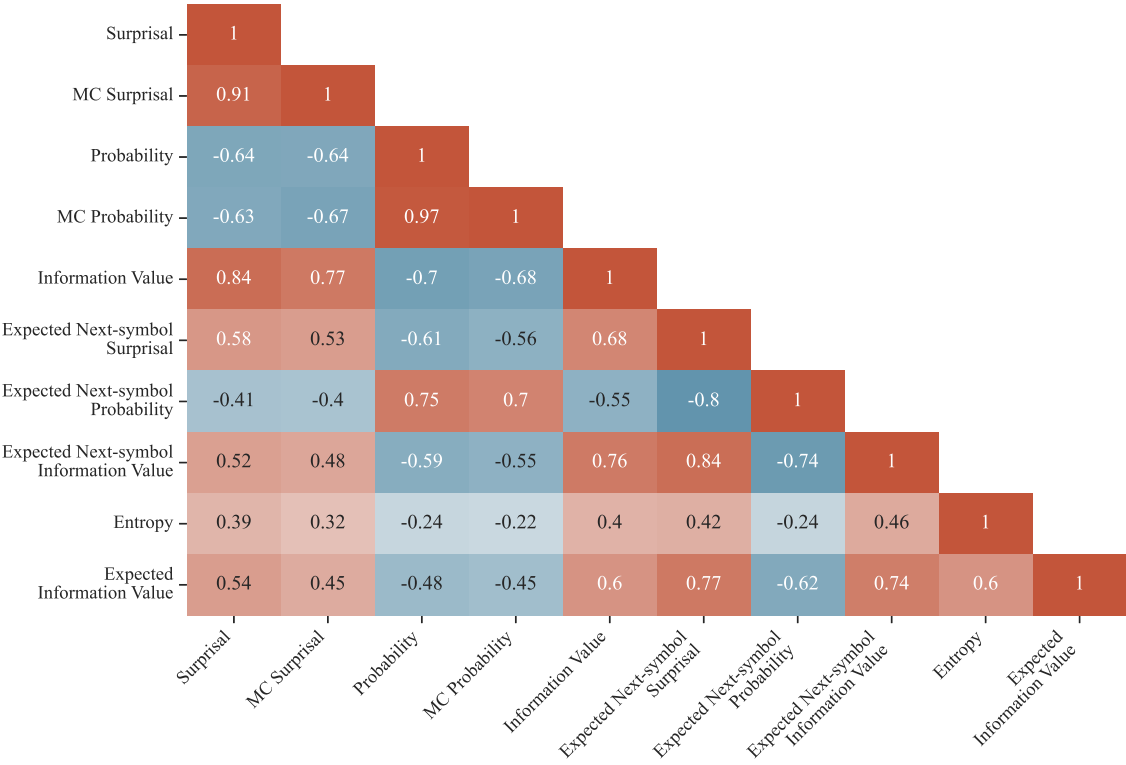
	Surprisal	MC Surprisal	Probability	MC Probability	Information Value	Expected Next-symbol Surprisal	Expected Next-symbol Probability	Expected Next-symbol Information Value	Entropy	Expected Information Value
Surprisal	1									
MC Surprisal	0.91	1								
Probability	-0.64	-0.64	1							
MC Probability	-0.63	-0.67	0.97	1						
Information Value	0.84	0.77	-0.7	-0.68	1					
Expected Next-symbol Surprisal	0.58	0.53	-0.61	-0.56	0.68	1				
Expected Next-symbol Probability	-0.41	-0.4	0.75	0.7	-0.55	-0.8	1			
Expected Next-symbol Information Value	0.52	0.48	-0.59	-0.55	0.76	0.84	-0.74	1		
Entropy	0.39	0.32	-0.24	-0.22	0.4	0.42	-0.24	0.46	1	
Expected Information Value	0.54	0.45	-0.48	-0.45	0.6	0.77	-0.62	0.74	0.6	1

Figure 5: Pearson correlation between responsive and anticipatory measures. Estimates obtained for the Aligned dataset. Monte Carlo (MC) samples with $N = 2^9$ and $L = 5$ from GPT-2 Small.
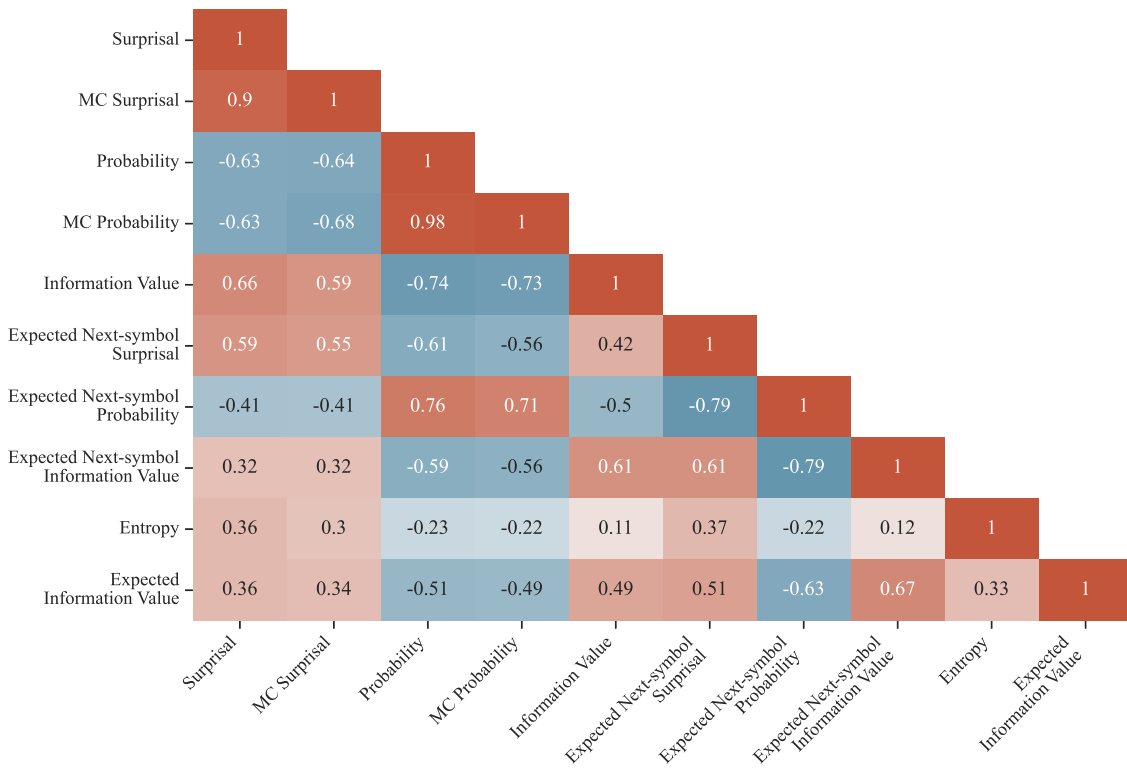
Figure 6: Pearson correlation between responsive and anticipatory measures. Estimates obtained for the Aligned dataset. Monte Carlo (MC) samples with $N = 2^9$ and $L = 5$ from GPT-Neo 125M.
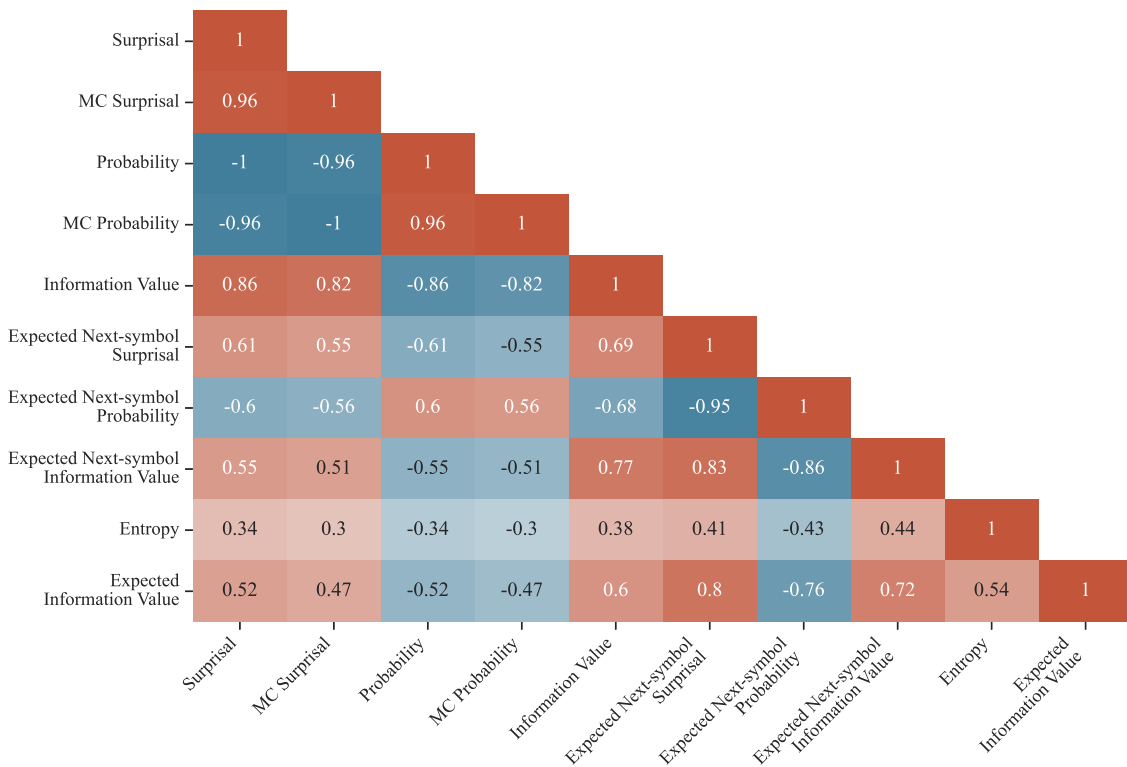

Figure 7: Spearman rank-correlation between responsive and anticipatory measures. Estimates obtained for the Aligned dataset. Monte Carlo (MC) samples with $N = 2^9$ and $L = 5$ from GPT-2 Small.
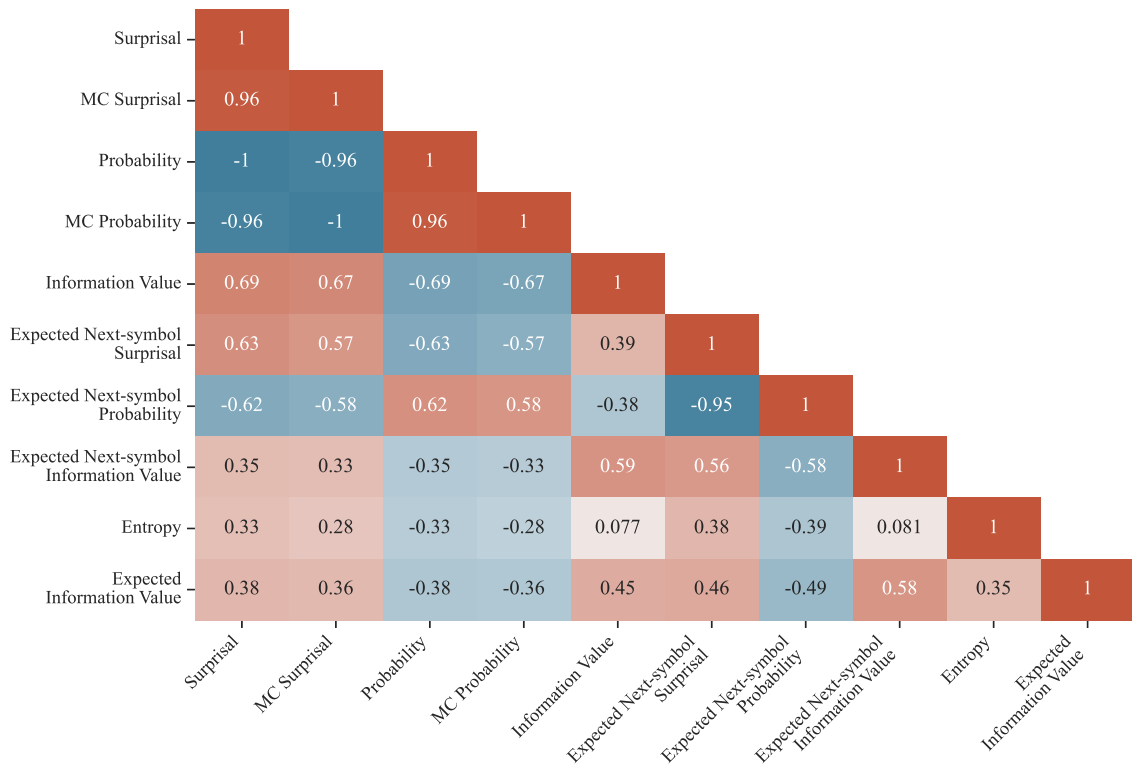
Figure 8: Spearman rank-correlation between responsive and anticipatory measures. Estimates obtained for the Aligned dataset. Monte Carlo (MC) samples with $N = 2^9$ and $L = 5$ from GPT-Neo 125M.
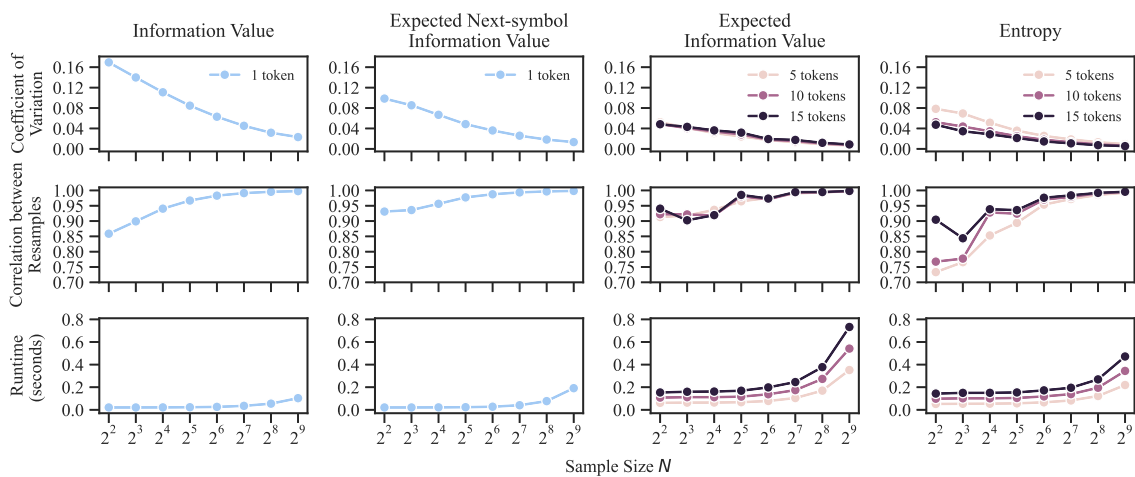
Figure 9: *Coefficient of variation* (top), *correlation between resamples* (center), and *runtimes* (bottom) for sampling-based measures across the stimuli in the Aligned dataset, using GPT-Neo 125M as a language model. Confidence intervals (95%) are too narrow to be visible; the horizontal axis is in log scale.
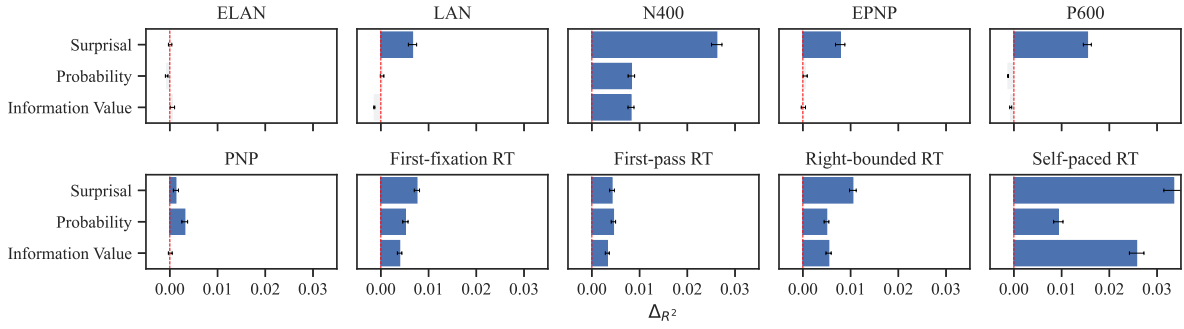
Figure 10: Predictive power Δ_{R^2} of responsive generalized surprisal models for event-related potentials and reading times, using GPT-Neo 125M as a language model; 95% confidence intervals. Significance color-coded: blue for $p < 0.0001$, gray for $p > 0.01$.
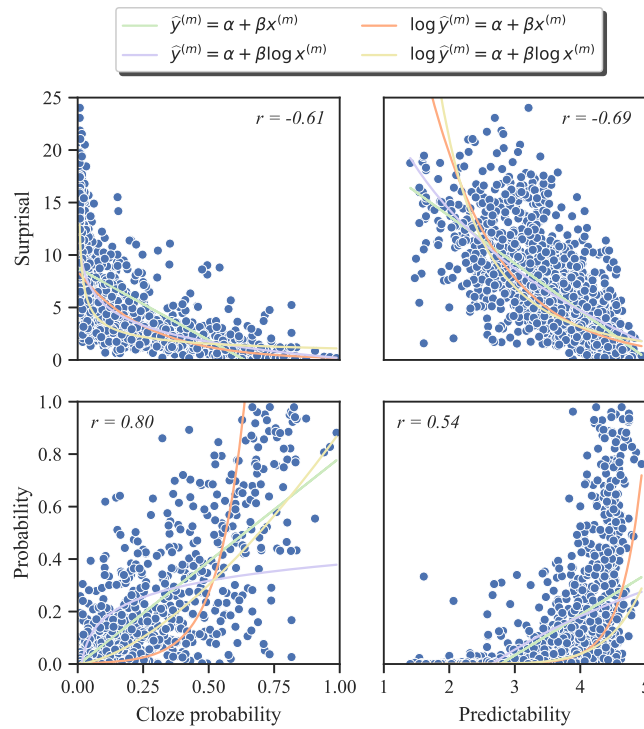
Figure 11: Probability and surprisal (GPT-Neo 125M) against human cloze probabilities and predictability ratings, with Pearson correlation coefficients r. For regressions, $x^{(m)} = \gamma(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})$ is the predictor and $y^{(m)} = \psi(\boldsymbol{w}^{(m)}, \boldsymbol{c}^{(m)})$ the predicted variable.
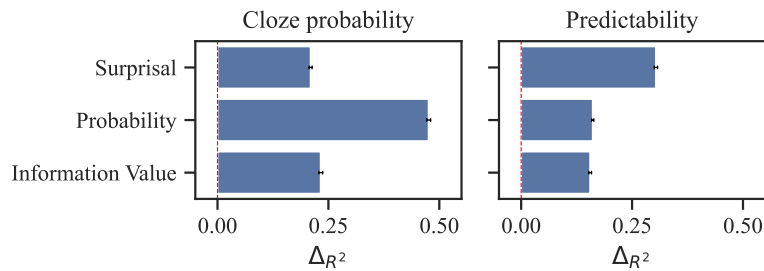
Figure 12: Predictive power Δ_{R^2} of responsive measures for human cloze probabilities and predictability ratings, using GPT-2 Small as a language model; 95% confidence intervals. The red dotted line represents the default baseline regressor. All measures are significantly predictive ($p < 0.001$).
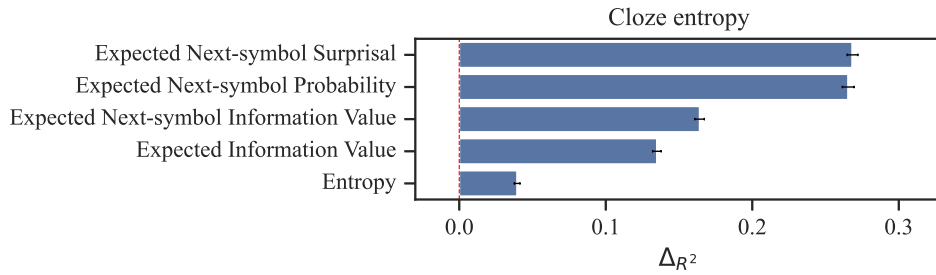
11667
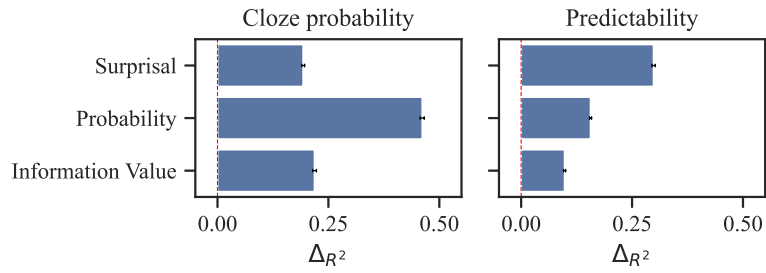
Figure 13: Predictive power Δ_{R^2} of responsive measures for human cloze probabilities and predictability ratings, using GPT-Neo 125M as a language model; 95% confidence intervals. The red dotted line represents the default baseline regressor. All measures are significantly predictive ($p < 0.001$).

Figure 14: Psycholinguistic predictive power Δ_{R^2} of anticipatory generalized surprisal models for the cloze entropy of the human cloze completion distributions, using GPT-2 Small as a language model; 95% confidence intervals. The red dotted line represents the default baseline regressor. All measures have significant predictive power ($p < 0.0001$).
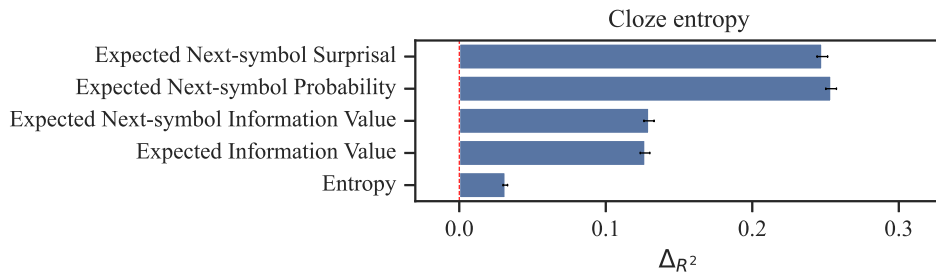

Figure 15: Psycholinguistic predictive power Δ_{R^2} of anticipatory generalized surprisal models for the cloze entropy of the human cloze completion distributions, using GPT-Neo 125M as a language model; 95% confidence intervals. The red dotted line represents the default baseline regressor. All measures have significant predictive power ($p < 0.0001$).
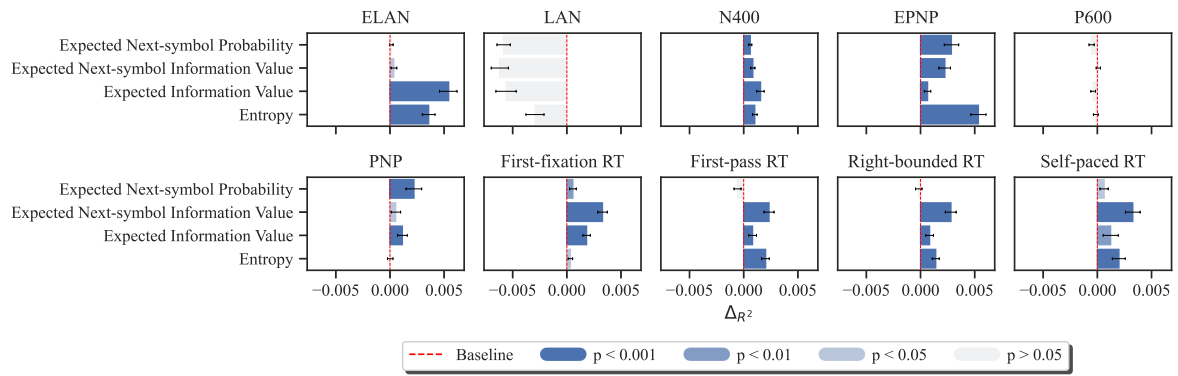
Figure 16: Difference between the predictive power of anticipatory measures used in combination with surprisal vs. expected next-symbol surprisal in combination with surprisal; 95% confidence intervals. GPT-Neo 125M is used as the language model. The red dotted line represents the combined baseline regressor. Significance is color-coded, as described in the legend.
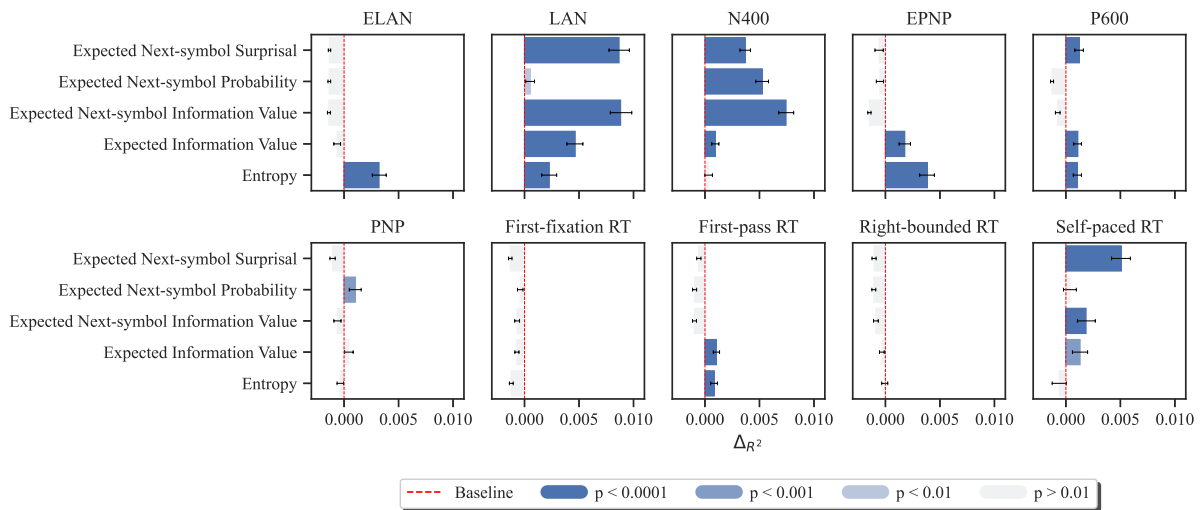

Figure 17: Psycholinguistic predictive power Δ_{R^2} of anticipatory generalized surprisal models for event-related potentials and reading times, using GPT-2 Small as a language model; 95% confidence intervals. The red dotted line represents the default baseline regressor. Significance is color-coded, as described in the legend.
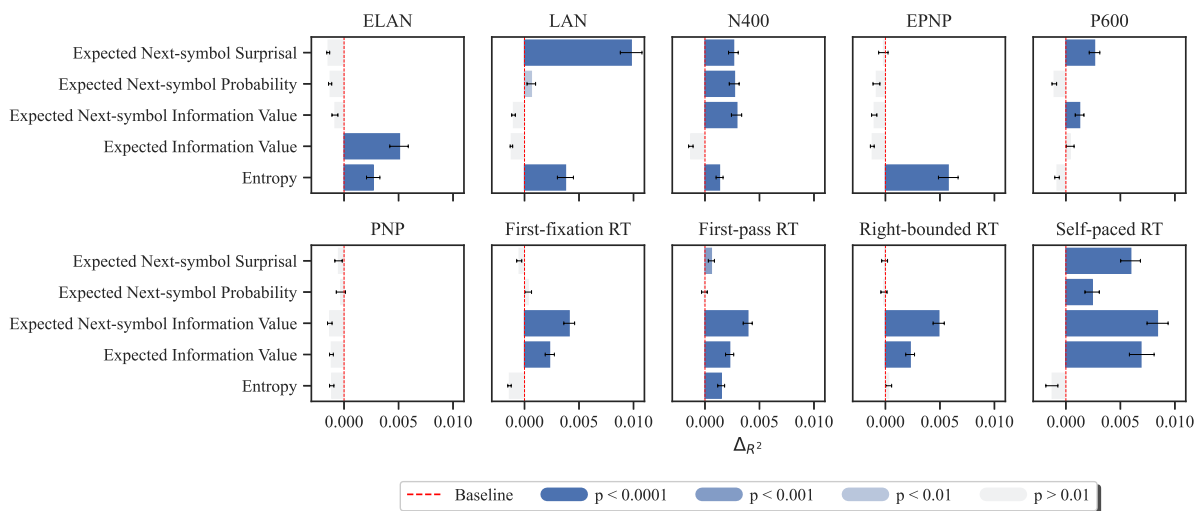

Figure 18: Psycholinguistic predictive power Δ_{R^2} of anticipatory generalized surprisal models for event-related potentials and reading times, using GPT-Neo 125M as a language model; 95% confidence intervals. The red dotted line represents the default baseline regressor. Significance is color-coded, as described in the legend.