Towards a Similarity-adjusted Surprisal Theory

Clara Meister **Mario Giulianelli** **Tiago Pimentel**
ETH Zürich, Department of Computer Science, Institute for Machine Learning
{clara.meister,mario.giulianelli,tiago.pimentel}@inf.ethz.ch
ETH zürich

Abstract

Surprisal theory posits that the cognitive effort required to comprehend a word is determined by its contextual predictability, quantified as surprisal. Traditionally, surprisal theory treats words as distinct entities, overlooking any potential similarity between them. Giulianelli et al. (2023) address this limitation by introducing information value, a measure of predictability designed to account for similarities between communicative units. Our work leverages Ricotta and Szeidl's (2006) diversity index to extend surprisal into a metric that we term similarity-adjusted surprisal, exposing a mathematical relationship between surprisal and information value. Similarity-adjusted surprisal aligns with information value when considering graded similarities and reduces to standard surprisal when words are treated as distinct. Experimental results with reading time data indicate that similarity-adjusted surprisal adds predictive power beyond standard surprisal for certain datasets, suggesting it serves as a complementary measure of comprehension effort.

1 Introduction

Surprisal theory (Hale, 2001) states that the effort a reader must spend to comprehend a word is a function of its contextual predictability, which is typically quantified as its surprisal, or negative log-probability. With numerous empirical (Smith and Levy, 2008, 2013; Shain, 2019, 2021, *inter alia*) and theoretical (Levy, 2008) studies supporting its claims, surprisal theory is a widely-accepted model of the effort required for sentence comprehension. Notably, surprisal theory treats words as completely distinct from one another, disregarding that they may express similar meanings. Motivated by this shortcoming, Giulianelli et al. (2023) proposed a new measure of comprehension effort: **information value**. Similarly to surprisal, information value quantifies the predictability of a linguistic unit in context; unlike surprisal, however, it accounts for communicative equivalences between possible continuations. Giulianelli et al. find this metric to be a significant predictor of utterance-level reading times and acceptability judgments, both independently and in addition to surprisal.

Similarly inspired, we investigate **similarity-adjusted surprisal** as a potential measure of comprehension effort. This measure is a natural extension of Ricotta and Szeidl's (2006) diversity index—which itself is a generalization of Shannon's entropy used to measure species biodiversity. Given a choice of similarity function, similarity-adjusted surprisal computes a word's predictability while considering its likeness to alternative continuations. Through this measure, we connect information value and standard surprisal, showing a mathematical relationship between these two metrics: similarity-adjusted surprisal has a monotonically increasing relationship with information value and reverts to standard surprisal when its similarity function regards different words as completely distinct.

In experiments with reading time data, we explore the psycholinguistic predictive power of similarity-adjusted surprisal with semantic, syntactic and orthographic notions of word distance. For some datasets, we see that—as with information value—similarity-adjusted surprisal provides significant predictive power above and beyond standard surprisal. This complementarity suggests that there are aspects of incremental comprehension effort that are not fully captured by the classic definitions used in surprisal theory. We also observe that non-contextual notions of semantic distance lead to better predictors than using contextual notions of distance, which supports observations that incremental comprehension effort is (at least partially) driven by shallow semantic processing.

2 Background

This section presents surprisal and information value, providing the relevant formal background

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16485–16498
November 12-16, 2024

for our similarity-adjusted surprisal. We will use $\mathcal{V}$ to refer to the vocabulary, i.e., a finite, non-empty set of words, and $\mathcal{V}^*$ to refer to the set of all strings formed by concatenating words in $\mathcal{V}$. We denote words as $w \in \mathcal{V}$, and strings (sequences of words) as $\boldsymbol{w} \in \mathcal{V}^*$. An index t, e.g., w_t and $\boldsymbol{w}_{<t}$, is used to mark positions within a string.

2.1 Surprisal Theory

According to surprisal theory, comprehending a word $w_t \in \mathcal{V}$ in its context $\boldsymbol{w}_{<t} \in \mathcal{V}^*$ requires a reader to update their beliefs about the intended meaning of the sentence, performing probabilistic inference over the space of possible meanings (Hale, 2001; Levy, 2008). The cost of this belief update is equal to the word's **surprisal**, or information content, whose formal definition is:[1]

$$\mathrm{h}(w_t) \stackrel{\text{def}}{=} -\log p(w_t \mid \boldsymbol{w}_{<t}) \tag{1}$$

If surprisal theory provides an accurate account of sentence comprehension, then we should find traces of this online inferential process in humans' behavioral responses to language comprehension tasks. In particular, assuming that a word's processing cost is reflected in its **reading time** (RT), a word's RT should be an increasing function of its surprisal (Smith and Levy, 2008). A large body of empirical work has examined the relationship between RT and surprisal, with results supporting surprisal theory (Smith and Levy, 2013; Wilcox et al., 2023; Shain et al., 2024, *inter alia*). Given these established results, new predictors of reading behavior, such as information value, would benefit from grounding in surprisal theory.

2.2 Information Value

Giulianelli et al. (2023) recently introduced a new measure to predict the cost associated with reading: **information value**. Let $\mathcal{A}_{\boldsymbol{w}_{<t}} \in \mathscr{P}(\mathcal{V}^*)$ be a multiset of plausible alternative continuations that a reader may expect to follow a given context $\boldsymbol{w}_{<t}$.[2] The information value of a continuation $\boldsymbol{w}_{\geq t} \in \mathcal{V}^*$ is defined as how different it is from continuations in $\mathcal{A}_{\boldsymbol{w}_{<t}}$. If $\boldsymbol{w}_{\geq t}$ is similar to what a reader expects, i.e., to elements of $\mathcal{A}_{\boldsymbol{w}_{<t}}$, then it does not convey much information, and should thus require little effort to process. If $\boldsymbol{w}_{\geq t}$ differs greatly from expected continuations, then it conveys more information and its processing cost is higher. Formally, we write $\boldsymbol{w}_{\geq t}$'s information value as $\boldsymbol{d}_{\boldsymbol{w}_{<t}}(\boldsymbol{w}_{\geq t}, \mathcal{A}_{\boldsymbol{w}_{<t}})$, where $\boldsymbol{d}_{\boldsymbol{w}_{<t}} : \mathcal{V}^* \times \mathscr{P}(\mathcal{V}^*) \to \mathbb{R}_+$ is a context-conditioned distance function.[3]

Following Giulianelli et al., we consider alternative sets $\mathcal{A}_{\boldsymbol{w}_{<t}}$ whose elements are sampled independently from $p(\cdot \mid \boldsymbol{w}_{<t})$ and distance functions $\boldsymbol{d}_{\boldsymbol{w}_{<t}}$ which apply element-wise to each instance in the alternative set through $d_{\boldsymbol{w}_{<t}} : \mathcal{V}^* \times \mathcal{V}^* \to \mathbb{R}_+$; we then aggregate individual distances by taking their mean. However, to make the comparison to standard next-word surprisal more natural, we take $\mathcal{A}_{\boldsymbol{w}_{<t}}$ to be composed of individual words rather than full string continuations. We will thus use the notation:

$$\boldsymbol{d}_{\boldsymbol{w}_{<t}}(w, \mathcal{A}_{\boldsymbol{w}_{<t}}) = \frac{\sum_{w' \in \mathcal{A}_{\boldsymbol{w}_{<t}}} d_{\boldsymbol{w}_{<t}}(w, w')}{|\mathcal{A}_{\boldsymbol{w}_{<t}}|} \tag{2}$$

This is a Monte Carlo estimator of the expected distance of a word w from other next words that start continuations of $\boldsymbol{w}_{<t}$ (Giulianelli et al., 2024b). We refer to it as **next-word information value**:

$$\mathrm{i}_d(w_t) \stackrel{\text{def}}{=} \sum_{w' \in \mathcal{V}} d_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid \boldsymbol{w}_{<t}) \tag{3}$$

3 Similarity-adjusted Surprisal Theory

To bridge the theoretical gap between surprisal and information value, we wish to derive a notion of per-word information content that accounts for similarities between different plausible continuations, rather than treating words as completely distinct outcomes. To this end, we turn to diversity indices: metrics developed in the field of biology to quantify biodiversity.[4] Analogous to our setting, when quantifying biodiversity, it is desirable to have a metric that takes into account that some species are more closely related to each other (e.g., two species in the same genus vs. in different genera). We adapt one of these metrics to our context.

3.1 Similarity-adjusted Entropy and Surprisal

Let R be a categorical random variable that takes on values $r \in \mathcal{R}$. Further, let $z : \mathcal{R} \times \mathcal{R} \to [0, 1]$ be a **similarity function**; it is 0 when r and r' are completely dissimilar and 1 when they are equivalent. Ricotta and Szeidl's (2006) diversity index is

[1] See App. B for a derivation and discussion of the relationship between $\mathrm{h}(w_t)$ and processing cost.

[2] $\mathscr{P}(\mathcal{V}^*)$ is the set of all multisets of elements in $\mathcal{V}^*$.

[3] This function may also be constant in $\boldsymbol{w}_{<t}$, for example, if $d_{\boldsymbol{w}_{<t}}$ measures orthographic distance between different $\boldsymbol{w}_{\geq t}$.

[4] For an overview, see Leinster and Cobbold (2012).

then defined as:

$$\mathrm{H}_z(R) \stackrel{\text{def}}{=} -\sum_{r\in\mathcal{R}} p(r)\log\sum_{r'\in\mathcal{R}} z(r,r')p(r') \quad (4)$$

If we use an identity similarity function such that $z(r,r')=1$ if $r=r'$ and 0 otherwise, then Eq. (4) is equivalent to Shannon's entropy (Shannon, 1948). We thus refer to Eq. (4) as **similarity-adjusted entropy**.

Bearing in mind entropy's close mathematical relationship to surprisal (i.e., entropy is the expected value of surprisal), similarity-adjusted entropy can be extended to a notion of surprisal that accounts for similarities between classes. We define the **similarity-adjusted surprisal** of outcome r as:

$$\mathrm{h}_z(r) \stackrel{\text{def}}{=} -\log\sum_{r'\in\mathcal{R}} z(r,r')\,p(r') \quad (5)$$

Comparably to Eq. (4), when the identity similarity function is used, then $\mathrm{h}_z(r) = -\log p(r) = \mathrm{h}(r)$. While a variant of similarity-adjusted surprisal has been used for measuring semantic uncertainty in neural machine translation (Cheng and Vlachos, 2024), its application in psycholinguistics has not yet been explored.

3.2 Similarity-adjusted Surprisal and Information Value

Now let $z_{\boldsymbol{w}_{<t}} : \mathcal{V}\times\mathcal{V}\to[0,1]$ be a similarity function that measures how similar two words are in context $\boldsymbol{w}_{<t}$. By using $z_{\boldsymbol{w}_{<t}}$ in Eq. (5), we arrive at a notion of similarity-adjusted surprisal for a word in context.

Definition 1. *The similarity-adjusted surprisal* h_z *of a word* w_t *in context* $\boldsymbol{w}_{<t}$ *is defined as:*

$$\mathrm{h}_z(w_t) \stackrel{\text{def}}{=} -\log\sum_{w'\in\mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t,w')\,p(w'\mid\boldsymbol{w}_{<t}) \quad (6)$$

Given this definition, we can now present the main theoretical result of this paper.

Theorem 1. *Let* $d_{\boldsymbol{w}_{<t}} : \mathcal{V}\times\mathcal{V}\to[0,1]$ *and* $z_{\boldsymbol{w}_{<t}}(w_t,w') = 1 - d_{\boldsymbol{w}_{<t}}(w_t,w')$. *Under these settings, next-word information value and similarity-adjusted surprisal have a monotonic, strictly increasing relationship.*

Proof. Proof in App. A. □

Note that this result trivially extends to information value and similarity-adjusted surprisal measured over finite strings $\boldsymbol{w}_{\geq t}\in\mathcal{V}^*$ of arbitrary length.

Because standard surprisal is a special case of similarity-adjusted surprisal, this result connects surprisal and information value; Giulianelli et al.'s findings can thus be seen as supporting an enriched notion of surprisal theory. App. B.2 shows how to use this relationship to derive a similarity-adjusted definition of processing cost.

3.3 Related Theories in Psycholinguistics

Several prior works in psycholinguistics have also examined variants of standard surprisal, e.g., decomposing (Roark et al., 2009; Li and Futrell, 2023), augmenting (Aurnhammer et al., 2021), or revising it (Arehalli et al., 2022; Giulianelli et al., 2023, 2024b,c). Some of these are motivated by the belief that the language comprehension process can be broken down into multiple distinct subtasks, for which there are different processing mechanisms (Kuperberg, 2016); they then associate variants of surprisal with these different cognitive processes. Roark et al. (2009), for instance, proposes that surprisal can be decomposed into a syntactic and a semantic component, each associated with its own cognitive process. In contrast to some of these works—for a subset of which we provide more detailed descriptions in App. C—we do not propose a decomposition of or alternative to surprisal theory. Rather, we view our work as offering a revised mathematical definition of surprisal, but still within the original surprisal theory framework.

4 Experimental Methodology

4.1 Data

We consider four datasets of naturalistic reading: Brown (Smith and Levy, 2013), Dundee (Kennedy et al., 2003), Natural Stories (Futrell et al., 2018), and Provo (Luke and Christianson, 2018). To collect these datasets, participants were administered text passages to read, and the time they spent fixating on each word was recorded. More details are provided in App. D. We organize these measurements into data points consisting of $\langle \mathbf{x}_n, y_n^{(i)}\rangle$ pairs, where $y_n^{(i)}\in\mathbb{R}_+$ is participant i's RT of word w_n, and $\mathbf{x}_n\in\mathbb{R}^d$ are word w_n's characteristics. These characteristics—which we refer to as predictors—consist of quantities such as a word's surprisal or unigram frequency. Following prior work (Wilcox et al., 2020; Meister et al., 2021, *inter alia*), we average RTs across participants, resulting in a single mean RT per word, $\bar{y}_n$. Our models are trained and tested to predict these averages.

4.2 Reading Time Regressors

Let f_{ψ} be a function that takes $\mathbf{x}_n$ and predicts $\bar{y}_n$. To avoid overlap in terminology with our discussion of language models, we refer to f_{ψ} as a regressor, and denote its parameters as ψ. A regressor f_{ψ} can take different functional forms. In light of prior work showing the surprisal–RT relationship to be largely linear (Smith and Levy, 2008, 2013; Wilcox et al., 2023; Shain et al., 2024), we restrict ourselves to linear f_{ψ}. Given a trained regressor f_{ψ} and a new data point $\mathbf{x}$, we can use f_{ψ} to either predict $\widehat{y} = f_{\psi}(\mathbf{x})$ or to estimate the probability of observing a specific $\bar{y}$ given an $\mathbf{x}$: $p_{\psi}(\bar{y} \mid \mathbf{x}) = \frac{(\bar{y} - f_{\psi}(\mathbf{x}))^2}{\sigma^2}$, where σ^2 is the regressor's variance estimated on the training set. The log-likelihood $\mathcal{L}(f_{\psi}, \mathcal{D})$ of a dataset $\mathcal{D}$ under f_{ψ} is then given by the (log of the) joint probability of observing those data points according to f_{ψ}.

4.3 Predictors

In all of our experiments, predictors $\mathbf{x}_n$ include two baseline variables typically used in RT analyses: word length (measured in characters) and unigram frequency. To account for spillover effects, which are caused by continued processing of previous words (Just et al., 1982; Frank et al., 2013), we include in $\mathbf{x}_n$ these variables for the current word w_n, as well as for the three words preceding it.

For our information-theoretic predictors, we use estimators. Our estimation of Eqs. (1), (2) and (6) can be summarized as: i) we replace the distribution p with a parameterized language model p_{θ}—specifically GPT-2 `small`—when computing Eqs. (1), (3) and (6);[5] ii) when it is too computationally expensive to sum over the entire vocabulary—which is required for exactly computing the expectations in Eqs. (3) and (6)—we use a Monte Carlo estimator (with 50 samples) for similarity-adjusted surprisal and next-word information value. More details on these information-theoretic estimators, as well as on methods for estimating unigram frequencies, are provided in App. D.

4.4 Similarity Functions

For both similarity-adjusted surprisal and information value, we consider several similarity functions; for each similarity function, we define an analogous distance as $d_{\boldsymbol{w}_{<t}}(w,w') = 1 - z_{\boldsymbol{w}_{<t}}(w,w')$. Details about precise estimation procedures are in App. D.

[5]This is standard practice in psycholinguistics (Smith and Levy, 2008; Goodkind and Bicknell, 2018; Wilcox et al., 2020, *inter alia*).

Word Embedding Similarity. Let $\phi : \mathcal{V} \rightarrow \mathbb{R}^d$ be a word embedding function, which may or not depend on context $\boldsymbol{w}_{<t}$. We compute the similarity between w and w' as the normalized cosine similarity:

$$z_{\boldsymbol{w}_{<t}}(w, w') = \frac{1}{2}\left(\frac{\phi(w) \cdot \phi(w')}{||\phi(w)||\,||\phi(w')||} + 1\right) \quad (7)$$

When computed using word embeddings, cosine similarity has proven itself a good metric of semantic similarity (Erk, 2009; Pennington et al., 2014). We again use GPT-2 `small` in all of our experiments as ϕ to produce non-contextual and contextual word embeddings.

Part-of-Speech Similarity. We use a measure of part-of-speech (POS) similarity as a notion of syntactic similarity:

$$z_{\boldsymbol{w}_{<t}}(w, w') = \begin{cases} 1, & \text{if } \textsc{pos}_{\boldsymbol{w}_{<t}}(w) = \textsc{pos}_{\boldsymbol{w}_{<t}}(w') \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where $\textsc{pos}_{\boldsymbol{w}_{<t}}$ is a POS-tagging model. We use the `pos-fast` model of the `flair` library.[6]

Orthographic Similarity. We further use a normalized version of string edit (Levenshtein) distance to quantify orthographic similarity. Let $d_{\text{L}}(w, w')$ be the edit distance between w and w'. Our orthographic similarity function is then

$$z_{\boldsymbol{w}_{<t}}(w, w') = 1 - \frac{d_{\text{L}}(w, w')}{\max\{|w|, |w'|\}} \quad (9)$$

where $|\cdot|$ measures string length in characters.

4.5 Evaluation

We quantify the predictive power of a predictor as the change in log-likelihood ($\Delta_{\mathcal{L}}$) of held-out data points $\mathcal{D}_{\text{test}}$ between a regressor f_{ψ_1} that includes this predictor and another, f_{ψ_2}, that does not:

$$\Delta_{\mathcal{L}} = \mathcal{L}(f_{\psi_1}, \mathcal{D}_{\text{test}}) - \mathcal{L}(f_{\psi_2}, \mathcal{D}_{\text{test}})\,. \quad (10)$$

This is a standard measure of predictive power in psycholinguistics (Goodkind and Bicknell, 2018; Wilcox et al., 2020). We estimate $\Delta_{\mathcal{L}}$ via 10-fold cross-validation: we use 9 data folds at a time to estimate the parameters of f_{ψ_1} and f_{ψ_2} and compute $\Delta_{\mathcal{L}}$ on the 10$^{\text{th}}$ fold; we report mean $\Delta_{\mathcal{L}}$ across folds. We run paired permutation tests with these 10-fold results to evaluate statistical significance.

[6]https://github.com/flairNLP/flair

	Similarity-adjusted surprisal				Information value			
	Non-contextual	Contextual	POS	Orthographic	Non-contextual	Contextual	POS	Orthographic
Brown	-0.02	-0.04	1.83	0.46	-0.01	-0.04	0.07	0.21
Dundee	0.13***	0.00	0.35	0.24	0.14***	0.00	0.01	0.02
Natural Stories	0.50***	0.32*	0.57	-0.03	0.58***	0.32*	0.04	0.07
Provo	-0.18***	0.04	0.86	-0.15	-0.19**	0.02	-0.01	-0.21

Table 1: $\Delta_{\mathcal{L}}$ (in 10^{-2} nats) over baseline *with* surprisal terms when adding a similarity-adjusted surprisal or information value term for each of the current and previous 3 words. Monte Carlo (MC) estimation with 50 samples.

5 Results and Discussion

Predictive power of similarity-adjusted surprisal. We evaluate the psycholinguistic predictive power that similarity-adjusted surprisal and information value provide beyond standard surprisal. To this end, we compute $\Delta_{\mathcal{L}}$ (Tab. 1) when adding these predictors to a regressor that already includes surprisal. In Natural Stories, we find similarity-adjusted surprisal and information value provide predictive power complementary to standard surprisal when using embedding-based (both contextual and non-contextual) similarity functions. In Dundee, significant additional predictive power only results from using non-contextual embedding similarities. Meanwhile, in Provo and Brown, similarity-adjusted surprisal and information value do not add predictive power beyond surprisal; they are significant predictors when evaluated against a control baseline, but less so than surprisal (see Tabs. 2 and 3 in App. E). In App. E, we also present results when exponentiating our definition of word pair similarity in Eq. (7) by an α such that similarity-adjusted surprisal converges to standard surprisal as $\alpha \to \infty$; we find that similarity-adjusted surprisal's predictive power is not strongly influenced by such choice.

The differences in predictive power of the two similarity-adjusted measures follow other notable trends across datasets. Predictive power is lowest on Provo, where stimuli have an average of only 50 words each, followed by Brown (553 words); it is highest for Natural Stories and Dundee, both containing stimuli whose average lengths are above 1000 words. These results suggest equipping surprisal with semantic measures of word predictability is beneficial when the psycholinguistic measurements at hand are collected for stimuli situated in broader discourse contexts. Other factors, such as the texts' style or topic, may have also played a role in these differences across datasets.

Broader implications. There is also an interpretation of these results as corroborating recently proposed theories of language comprehension. The Natural Stories corpus contains low-frequency (albeit still grammatically correct) syntactic constructions. Thus, in this corpus, we encounter continuations that are less predictable from the context but have high similarity with more predictable alternative continuations. The result that our semantic variants of similarity-adjusted surprisal and information value provide significant predictive power over standard surprisal, particularly for this dataset, can be taken as support for models of heuristic processing (e.g., Li and Futrell, 2023; see App. C for further discussion). Similarly, the overall higher predictive power provided by non-contextual similarity functions (when compared to contextual ones) could be taken as evidence that incremental comprehension effort is more sensitive to forms of shallow semantic processing (Barton and Sanford, 1993; Daneman et al., 2007) than to deep integration of contextualized word meaning. However, due to the known sensitivity of contextual embeddings to word-unspecific sentential information (Klafka and Ettinger, 2020; Erk and Chronis, 2022), further analysis with semantic similarity functions is required to confirm this finding.

6 Conclusion

This work introduces similarity-adjusted surprisal: a measure of contextual word predictability that takes into account word similarities. By equipping surprisal with the ability to account for words' relationships, we reconcile surprisal theory's predictions with those of information value and demonstrate their mathematical relationship. Our experimental results on RT data indicate similarity-adjusted surprisal has predictive power beyond that of standard surprisal, thus validating and enriching surprisal theory. Points for future research include analyzing similarity functions that capture different characteristics of word meaning, as well as measuring the predictive power of similarity-adjusted surprisal for other indices of processing difficulty, such as N400 and other event-related brain potentials.

Limitations

We do not provide a comprehensive assessment of different design choices and experimental settings, limiting the definitiveness with which we can draw conclusions about the efficacy of similarity-adjusted surprisal as a predictor of language comprehension. These different choices and settings deserve further exploration. The reading time datasets that we employ are in English, and thus, we can only draw conclusions about reading behavior in the English language. Further, we only consider two functions for computing word similarities. As the functional form of the surprisal–reading time relationship has proven to be quite important for the psycholinguistic predictive power of surprisal, it is conceivable that the choice of similarity function could likewise have a large impact on the psycholinguistic predictive power of our diverse predictors.

Acknowledgments

Clara Meister was supported by a Google PhD Fellowship. Mario Giulianelli was supported by an ETH Zurich Postdoctoral Fellowship. We thank our anonymous reviewers for their insightful feedback and helpful pointers to related works.

References

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.

Christoph Aurnhammer, Francesca Delogu, Miriam Schulz, Harm Brouwer, and Matthew W. Crocker. 2021. Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, 16(9):1–31.

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

Julius Cheng and Andreas Vlachos. 2024. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128, St. Julian's, Malta. Association for Computational Linguistics.

Meredyth Daneman, Tracy Lennertz, and Brenda Hannon. 2007. Shallow semantic processing of text: Evidence from eye movements. *Language and Cognitive Processes*, 22(1):83–105.

Katrin Erk. 2009. Representing words as regions in vector space. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 57–65, Boulder, Colorado. Association for Computational Linguistics.

Katrin Erk and Gabriella Chronis. 2022. Word embeddings are word story embeddings (and that's fine). In *Algebraic Structures in Natural Language*, pages 189–218. CRC Press.

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

Stefan L. Frank and Roel M. Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

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*. European Language Resources Association.

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, Andreas Opedal, and Ryan Cotterell. 2024b. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA. Association for Computational Linguistics.

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

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.

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*, pages 1–8.

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.

Alan Kennedy, Robin Hill, and Joel Pynte. 2003. The Dundee corpus. In *Proceedings of the 12th European Conference on Eye Movements*.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Gina Kuperberg. 2016. Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 31:1–15.

Tom Leinster and Christina A Cobbold. 2012. Measuring diversity: The importance of species similarity. *Ecology*, 93(3):477–489.

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

Jixuan Li and Richard Futrell. 2023. A decomposition of surprisal tracks the N400 and P600 brain potentials. volume 45. Proceedings of the Annual Meeting of the Cognitive Science Society.

Steven G. Luke and Kiel Christianson. 2018. The Provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Irene Nikkarinen, Tiago Pimentel, Damián Blasi, and Ryan Cotterell. 2021. Modeling the unigram distribution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3721–3729, Online. Association for Computational Linguistics.

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.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary poses a confound for calculating word probabilities. *Preprint*, arXiv:2406.10851.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*, 70(3):237–243.

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.

Douglas Roland, Hongoak Yun, Jean-Pierre Koenig, and Gail Mauner. 2012. Semantic similarity, predictability, and models of sentence processing. *Cognition*, 122(3):267–279.

Cory Shain. 2019. A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4086–4094, Minneapolis, Minnesota. Association for Computational Linguistics.

Cory Shain. 2021. CDRNN: Discovering complex dynamics in human language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3718–3734, Online. Association for Computational Linguistics.

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.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Nathaniel J. Smith and Roger Levy. 2008. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings 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.

Mark Steedman. 1987. Combinatory grammars and parasitic gaps. *Natural Language & Linguistic Theory*, 5(3):403–439.

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

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Cognitive Science Society*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Relationship between Information Value and Similarity-adjusted Surprisal

Theorem 1. *Let* $d_{\boldsymbol{w}_{<t}} : \mathcal{V} \times \mathcal{V} \to [0, 1]$ *and* $z_{\boldsymbol{w}_{<t}}(w_t, w') = 1 - d_{\boldsymbol{w}_{<t}}(w_t, w')$. *Under these settings, next-word information value and similarity-adjusted surprisal have a monotonic, strictly increasing relationship.*

$$\mathrm{i}_d(w_t) \overset{\propto}{\nearrow} \mathrm{h}_z(w_t) \tag{11}$$

Proof. Using the relationship $d_{\boldsymbol{w}_{<t}}(w_t, w') = 1 - z_{\boldsymbol{w}_{<t}}(w_t, w')$, simple algebraic manipulation shows that:

$$\begin{aligned}
\mathrm{i}_d(w_t) &= \sum_{w' \in \mathcal{V}} p(w' \mid \boldsymbol{w}_{<t})\, d_{\boldsymbol{w}_{<t}}(w_t, w') \\
&= \sum_{w' \in \mathcal{V}} p(w' \mid \boldsymbol{w}_{<t})\, \left(1 - z_{\boldsymbol{w}_{<t}}(w_t, w')\right) \\
&= 1 - \sum_{w' \in \mathcal{V}} p(w' \mid \boldsymbol{w}_{<t})\, z_{\boldsymbol{w}_{<t}}(w_t, w') \\
&\overset{\propto}{\nearrow} -\log \sum_{w' \in \mathcal{V}} p(w' \mid \boldsymbol{w}_{<t})\, z_{\boldsymbol{w}_{<t}}(w_t, w') \\
&= \mathrm{h}_z(w_t \mid \boldsymbol{w}_{<t})
\end{aligned} \tag{12}$$

where $\overset{\propto}{\nearrow}$ indicates a monotonic, strictly increasing relationship and follows from the fact that $\log$ is a monotone, strictly increasing function. □

B The Mathematical Relationship between Processing Costs and Surprisal

B.1 Equivalence between Surprisal and KL Divergence

A word's surprisal is equivalent to the Kullback–Leibler (KL) divergence between two probability distributions over a sentence's potential meanings: one with and one without knowledge of that word (Levy, 2008). Formally, let $\mathcal{M}$ be the space of potential sentence meanings, let $m \in \mathcal{M}$ be a meaning and let $p(m \mid \boldsymbol{w}_{<t})$ be the probability of meaning $m \in \mathcal{M}$ conditioned on a prefix $\boldsymbol{w}_{<t}$. We can define the cost of reading a word w_t as the amount of energy spent to update this distribution. Assuming this energy consumption is a function of the distance between the prior and posterior distributions over meanings after observing w_t, we can define cost according to the KL divergence between these two distributions, a standard measure of the difference between distributions:

$$\begin{aligned}
\mathrm{cost}(w_t) &\overset{\text{def}}{=} \mathrm{KL}(p(m \mid \boldsymbol{w}_{<t} \circ w_t) \,||\, p(m \mid \boldsymbol{w}_{<t})) \\
&= \sum_{m \in \mathcal{M}} p(m, \boldsymbol{w}_{<t} \circ w_t) \log \frac{p(m \mid \boldsymbol{w}_{<t} \circ w_t)}{p(m \mid \boldsymbol{w}_{<t})} \,.
\end{aligned} \tag{13}$$

Under standard assumptions,[7] we can show that this divergence is equivalent to word w_t's surprisal.

Theorem 2. ***Cost equals information content*** *(result from Levy, 2008, reiterated here in the notation of this paper). Under standard assumptions about* $p(m, \boldsymbol{w}_{<t} \circ w_t)$*, we can show that:*

$$\mathrm{cost}(w_t) = \mathrm{h}(w_t) \,. \tag{14}$$

Proof. Let $p(\boldsymbol{w}_{<t} \circ w_t \mid m)$ be deterministic, i.e., there is only one sequence $\boldsymbol{w}_{<t} \circ w_t$ which can be used

[7] We assume, as in Levy (2008), that $p(\boldsymbol{w}_{<t} \circ w_t \mid m)$ is deterministic, i.e., there is only one sequence $\boldsymbol{w}_{<t} \circ w_t$ which can be used to convey each meaning m. While perhaps unrealistic, we can still draw insights from this result.

to convey each meaning m. From Bayes theorem, we then have that:

$$p(m \mid \boldsymbol{w}_{<t} \circ w_t) = \frac{\overbrace{p(w_t \mid m, \boldsymbol{w}_{<t})}^{=1 \text{ because deterministic}} \; p(m \mid \boldsymbol{w}_{<t})}{p(w_t \mid \boldsymbol{w}_{<t})} \tag{15a}$$

$$= \frac{p(m \mid \boldsymbol{w}_{<t})}{p(w_t \mid \boldsymbol{w}_{<t})} \tag{15b}$$

We now use this equality to arrive at the desired result:

$$\begin{aligned}
\text{cost}(w_t) &= \text{KL}(p(m \mid \boldsymbol{w}_{<t} \circ w_t) \mid\mid p(m \mid \boldsymbol{w}_{<t})) && \text{(16a)}\\
&= \sum_{m \in \mathcal{M}} p(m, \boldsymbol{w}_{<t} \circ w_t) \log \frac{p(m \mid \boldsymbol{w}_{<t} \circ w_t)}{p(m \mid \boldsymbol{w}_{<t})} && \text{(16b)}\\
&= \sum_{m \in \mathcal{M}} p(m, \boldsymbol{w}_{<t} \circ w_t) \log \frac{p(m \mid \boldsymbol{w}_{<t})}{p(w_t \mid \boldsymbol{w}_{<t})\, p(m \mid \boldsymbol{w}_{<t})} && \text{(16c)}\\
&= \sum_{m \in \mathcal{M}} p(m, \boldsymbol{w}_{<t} \circ w_t) \log \frac{1}{p(w_t \mid \boldsymbol{w}_{<t})} && \text{(16d)}\\
&= \log \frac{1}{p(w_t \mid \boldsymbol{w}_{<t})} && \text{(16e)}\\
&= \text{h}(w_t) && \text{(16f)}
\end{aligned}$$

□

B.2 Similarity-adjusted Surprisal as a KL Divergence

In this section, we first define a number of similarity-aware distributions as:

$$p_z(w_t \mid \boldsymbol{w}_{<t}) \stackrel{\text{def}}{=} \sum_{w' \in \mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid \boldsymbol{w}_{<t}) \qquad \text{expectation in similarity-adjusted surprisal} \tag{17a}$$

$$p_z(w_t \mid m, \boldsymbol{w}_{<t}) \stackrel{\text{def}}{=} \sum_{w' \in \mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid m, \boldsymbol{w}_{<t}) \qquad \text{analogous to } p_z(w_t \mid \boldsymbol{w}_{<t}) \tag{17b}$$

$$p_z(m \mid \boldsymbol{w}_{<t}) \stackrel{\text{def}}{=} p(m \mid \boldsymbol{w}_{<t}) \qquad \text{does not depend on } w_t \tag{17c}$$

$$p_z(m \mid \boldsymbol{w}_{<t} \circ w_t) \stackrel{\text{def}}{=} \frac{p_z(w_t \mid m, \boldsymbol{w}_{<t})\, p_z(m \mid \boldsymbol{w}_{<t})}{p_z(w_t \mid \boldsymbol{w}_{<t})} \qquad \text{Bayes-inspired definition} \tag{17d}$$

where Eq. (17d) is "Bayes-inspired" because p_z is not necessarily a valid probability distribution (it does not necessarily sum to 1 across its support) and so the equivalence given by Bayes theorem is not guaranteed; rather it is an equivalence that we enforce in the definition of p_z.

We now present a theorem linking similarity-adjusted surprisal and processing cost, under the definitions above.

Theorem 3. ***Cost equals similarity-adjusted surprisal.*** *Under standard assumptions*[7] *about* $p_z(m, \boldsymbol{w}_{<t} \circ w_t)$ *and using the definitions in Eq.* (17)*, we can show that:*

$$\text{cost}_z(w_t) = \text{h}_z(w_t) \tag{18}$$

Proof. First, we provide a helpful equivalence for $p_z(m \mid \boldsymbol{w}_{<t} \circ w_t)$:

$$p_z(m \mid \boldsymbol{w}_{<t} \circ w') = \frac{p_z(w' \mid m, \boldsymbol{w}_{<t})\, p_z(m \mid \boldsymbol{w}_{<t})}{p_z(w' \mid \boldsymbol{w}_{<t})} \tag{19a}$$

$$= \frac{\left(\sum_{w' \in \mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid m, \boldsymbol{w}_{<t})\right) p(m \mid \boldsymbol{w}_{<t})}{\sum_{w' \in \mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid \boldsymbol{w}_{<t})} \qquad \text{expand } p_z \tag{19b}$$

$$= \frac{z_{\boldsymbol{w}_{<t}}(w_t, w_t) p(w_t \mid m, \boldsymbol{w}_{<t})\, p(m \mid \boldsymbol{w}_{<t})}{\sum_{w' \in \mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid \boldsymbol{w}_{<t})} \qquad \text{deterministic } p(w' \mid m, \boldsymbol{w}_{<t}) \tag{19c}$$

$$= \frac{p(m \mid \boldsymbol{w}_{<t})}{\sum_{w' \in \mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid \boldsymbol{w}_{<t})} \qquad p(w_t \mid m, \boldsymbol{w}_{<t}) = 1 \tag{19d}$$

Given these equalities, we can follow the same logic as in Theorem 2 to show that processing cost in the presence of similarity-aware distributions over words has an equivalence with similarity-adjusted surprisal:

$$\text{cost}_z(w_t \mid \boldsymbol{w}_{<t}) = \text{KL}_z(p(m \mid \boldsymbol{w}_{<t} \circ w_t) \mid\mid p(m \mid \boldsymbol{w}_{<t})) \tag{20a}$$

$$= \sum_{m \in \mathcal{M}} p(m \mid \boldsymbol{w}_{<t} \circ w_t) \log \frac{p_z(m \mid \boldsymbol{w}_{<t} \circ w_t)}{p_z(m \mid \boldsymbol{w}_{<t})} \tag{20b}$$

$$= \sum_{m \in \mathcal{M}} p(m \mid \boldsymbol{w}_{<t} \circ w_t) \log \frac{p(m \mid \boldsymbol{w}_{<t})}{\left(\sum_{w' \in \mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid \boldsymbol{w}_{<t})\right) p(m \mid \boldsymbol{w}_{<t})} \tag{20c}$$

$$= \sum_{m \in \mathcal{M}} p(m \mid \boldsymbol{w}_{<t} \circ w_t) \log \frac{1}{\sum_{w' \in \mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid \boldsymbol{w}_{<t})} \tag{20d}$$

$$= \log \frac{1}{\sum_{w' \in \mathcal{V}} z_{\boldsymbol{w}_{<t}}(w_t, w')\, p(w' \mid \boldsymbol{w}_{<t})} \tag{20e}$$

$$= -\log p_z(w_t \mid \boldsymbol{w}_{<t}) \tag{20f}$$

$$= \text{h}_z(w_t) \tag{20g}$$

where KL_z is defined analogously to both standard KL and similarity-adjusted entropy: it implements the same expectation as standard KL, but takes the log of distributions p_z instead. □

C Related Work in Psycholinguistics

In this section, we review some related work in more detail, and when possible connect it to similarity-adjusted surprisal. Arehalli et al. (2022) investigate a word's syntactic surprisal—i.e., the surprisal associated with the syntactic structure implied by that word—as a predictor of reading comprehension behavior. They define this value as:

$$-\log \sum_{w' \in \mathcal{V}} \underbrace{\sum_{\text{POS} \in \mathcal{C}} p(\text{POS} \mid \boldsymbol{w}_{<t} \circ w_t)\, p(\text{POS} \mid \boldsymbol{w}_{<t} \circ w')}_{\textit{potential choice of } z_{\boldsymbol{w}_{<t}}(w_t, w')}\; p(w' \mid \boldsymbol{w}_{<t}) \tag{21}$$

where, in their case, $\text{POS} \in \mathcal{C}$ represents a combinatory categorial grammar (CCG Steedman, 1987) supertag. We provide a more general measure of predictability here, as we can realize their notion of syntactic surprisal in our similarity-adjusted surprisal framework by using a similarity function that identifies equivalent syntactic classes.

More broadly speaking, many works have employed notions of semantic similarity in their models of language comprehension (Roland et al., 2012; Frank and Willems, 2017; Giulianelli et al., 2023, 2024b,c; Li and Futrell, 2023, *inter alia*). For example, Li and Futrell (2023) offers a decomposition of a word's surprisal into two quantities, where they specifically consider the word as perceived by a comprehender ω_t:

$$\text{h}(\omega_t) \overset{\text{def}}{=} -\log p(\omega_t \mid \boldsymbol{\omega}_{<t}) = \underbrace{\mathbb{E}\left[-\log p(w_t \mid \boldsymbol{w}_{<t})\right]}_{\textit{heuristic surprise}} + \underbrace{\mathbb{E}\left[\log \frac{p(w_t \mid \boldsymbol{w}_{<t})}{p(\omega_t \mid \boldsymbol{\omega}_{<t})}\right]}_{\textit{discrepancy signal}} \tag{22}$$

Here, w_t represents the ground truth word at time t, which they call a "heuristic word". Similarly to our semantic variant of similarity-adjusted surprisal, they use a notion of semantic distance to estimate the latter quantity. Our works differ in several ways though, most notably in that we do not propose a new model of language comprehension, but rather introduce an alternative definition of surprisal.

D Experimental Setup

Code for reproducing experimental results can be found at https://github.com/cimeister/diverse-surprisal.

D.1 Data

We use four reading time datasets. Per-word reading time is measured according to one of two paradigms: self-paced and eye-tracked reading. The self-paced reading corpora are the Natural Stories Corpus (Futrell et al., 2018) and the Brown Corpus (Smith and Levy, 2013). The eye-tracking corpora are the Provo Corpus (Luke and Christianson, 2018) and the Dundee Corpus (Kennedy et al., 2003). We refer to the original works for further details on data collection.

Before computing our different word-level predictors, text from all corpora was pre-processed using the Moses decoder[8] tokenizer and punctuation normalizer. Additional pre-processing was performed by the tokenizers for respective neural models. Capitalization was kept intact albeit we used the lowercase version of words when querying for unigram frequency estimates. We estimate unigram frequencies following Nikkarinen et al. (2021) on the WikiText 103 dataset.

D.2 Information-Theoretic Estimators

We estimate surprisal and information value[9] using GPT-2 `small` (Radford et al., 2019);[10] while not the most accurate language model in terms of perplexity, prior work has shown GPT-2 `small` to have better psycholinguistic predictive power than its larger counterparts (Oh and Schuler, 2023; Shain et al., 2024). Note that GPT-2 `small` operates over subwords while reading time measurements are taken at the word level. We discuss our approaches for accounting for this characteristic for each estimator separately.

Surprisal. We query our language model for next token probabilities; our surprisal estimate for a token is then simply the negative log of this value. To compute word-level estimates of surprisal, we sum these values across the tokens that constitute each word (as delineated by the reading time dataset). In general, surprisal decomposes additively across subunits of a word, theoretically grounding this approach. However, as Pimentel and Meister (2024) point out, the way that subword units demarcate the beginning of a word complicates the computation of word-level surprisal estimates. They offer a simple fix for this issue, which we do not incorporate here since extending it to information value and similarity-adjusted surprisal is non-trivial. See also Oh and Schuler (2024) for a similar discussion, and Giulianelli et al. (2024a) for further discussion on the role of tokenization in computational psycholinguistics, as well as for a method to compute the surprisal of any character span from subword-level language models.

Embedding-based information value and similarity-adjusted surprisal estimators. We likewise use GPT-2 `small` for our word embeddings. Transformer-based language models can provide word embeddings for all of the (sub)words in their vocabulary. Thus, in this setting, we take $\mathcal{V}$ to be GPT-2 `small`'s subword vocabulary. We explore both contextual and non-contextual word embeddings in the computation of Eq. (7). We use layer 0 for non-contextual and layer 12 (the last layer) for contextual embeddings; we leave the exploration of the use of other embedding functions, e.g., other language models, layers or aggregation across layers, to future work. In the case of contextual embeddings, obtaining embeddings for each $w \in \mathcal{V}$ requires a separate query to the language model. Querying the model $|\mathcal{V}| \approx 50,000$ times for every context $\boldsymbol{w}_{<t}$ would be very computationally intensive, so we instead use a Monte Carlo estimator for these variants of information value and similarity-adjusted surprisal.

[8] http://www.statmt.org/moses/

[9] We use the codebase of Giulianelli et al. (2023) to compute information value. For variants of similarity-adjusted surprisal and information value that require estimators, we use 50 samples in all experiments.

[10] We use the open-source version available on the `transformers` library (Wolf et al., 2020).

Specifically, similarly to Giulianelli et al. (2024b), we sample next tokens (with replacement) according to $p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{w}_{<t})$. We query our language model for only the embeddings for these tokens, and use them to make a Monte Carlo estimator of Eqs. (2) and (6).[11] To create next-word information value and similarity-adjusted surprisal estimates from these subword-level estimates, we sum across these values for each of the tokens that constitute a word. Note that information value and similarity-adjusted surprisal do not decompose across subwords. We ran experiments where predictors were set to the value of the first subword that constituted a word and observed similar results; we omit these to reduce clutter.

POS and edit distance estimators. These estimators cannot be computed at the subword level. Thus, we consider a full-word vocabulary. Because of the computational load that it would require to get language model estimates for a comprehensive vocabulary, we instead use a Monte Carlo estimator. Explicitly, we sample full-word continuations (with replacement) according to $p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{w}_{<t})$, sampling subwords until we reach either a white space marker or the end-of-sentence token. Note that this also allows us to avoid the task of explicitly defining a full-word vocabulary. We then use these continuations to build Monte Carlo estimators of Eqs. (2) and (6).

E Additional Experimental Results

E.1 From Similarity-adjusted to Standard Surprisal

In this experiment we equip $z_{\boldsymbol{w}_{<t}}$ with a temperature parameter α, exponentiating our definition of word pair similarity in Eq. (7) by a given α. As $\alpha \to \infty$, all $z_{\boldsymbol{w}_{<t}}(w, w')$ for $w \neq w'$ go to 0; on the other hand $z_{\boldsymbol{w}_{<t}}(w, w)$ remains at 1. Thus, similarity-adjusted surprisal converges to standard surprisal as $\alpha \to \infty$. We observe how the psycholinguistic predictive power of similarity-adjusted surprisal changes with α in Fig. 1. While varying α does not appear to have a significant effect on the predictive power of similarity-adjusted surprisal, we see that, as expected, the $\Delta_{\mathcal{L}}$ of h_{z^α} converges to that of surprisal as $\alpha \to \infty$.

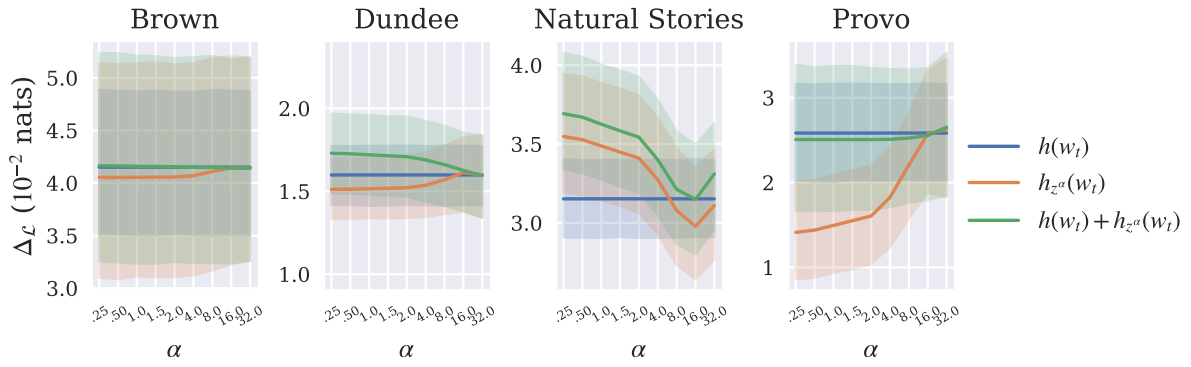

Figure 1: The change in reading time dataset log-likelihoods as a function of the temperature parameter used with the semantic-similarity function in similarity-adjusted surprisal computations. Each line corresponds to a different set of predictors added to the regressor. Shaded regions indicate 95% confidence intervals, as computed using standard bootstrapping techniques on our per-fold $\Delta_{\mathcal{L}}$ values.

[11] For non-contextual embeddings, we compute Eqs. (2) and (6) exactly, with GPT-2 `small`'s vocabulary as $\mathcal{V}$.

E.2 $\Delta_{\mathcal{L}}$ of Different Predictors

	Surprisal	**Similarity-adjusted surprisal**				**Information value**			
		Non-contextual	Contextual	POS	Orthographic	Non-contextual	Contextual	POS	Orthographic
Brown	4.15***	2.78***	-0.02	2.22	1.29*	2.61***	-0.02	0.33*	0.71*
Dundee	1.60***	1.44***	0.01	0.69	0.64*	1.43***	0.01	0.26***	0.48***
Natural Stories	3.15***	3.18***	0.31**	1.04*	0.79*	3.18***	0.30***	0.46***	0.88***
Provo	2.58***	1.34***	0.05	1.49	0.82	1.14***	0.06	0.57	0.83

Table 2: $\Delta_{\mathcal{L}}$ (in 10^{-2} nats) on reading time data of regressors with different predictors over baseline regressors (i.e., regressors with only baseline predictors). Variable values for current and previous 3 words are provided as predictors. MC estimates of information value and similarity-adjusted surprisal use 50 samples per context.

	Similarity-adjusted surprisal				**Information value**			
	Non-contextual	Contextual	POS	Orthographic	Non-contextual	Contextual	POS	Orthographic
Brown	-1.37***	-4.17***	-1.94	-2.86***	-1.54***	-4.17***	-3.82***	-3.44***
Dundee	-0.16*	-1.59***	-0.91	-0.96***	-0.17*	-1.59***	-1.34***	-1.12***
Natural Stories	0.03	-2.84***	-2.11**	-2.37***	0.03	-2.85***	-2.69***	-2.27***
Provo	-1.24**	-2.54***	-1.09	-1.76*	-1.44***	-2.52***	-2.02**	-1.75**

Table 3: $\Delta_{\mathcal{L}}$ (in 10^{-2} nats) on reading time data of regressors with our different predictors of interest in comparison to regressors *with* surprisal (i.e., replacing all surprisal terms for current and previous words with corresponding similarity-adjusted surprisal/information value terms). MC estimates of information value and similarity-adjusted surprisal use 50 samples per context.

	Similarity-adjusted surprisal				**Information value**			
	Non-contextual	Contextual	POS	Orthographic	Non-contextual	Contextual	POS	Orthographic
Brown	0.01	-0.01	1.00	-0.03	0.02	-0.01	0.00	0.01
Dundee	0.12***	-0.00	0.38	0.22**	0.14***	-0.00	-0.00	0.00
Natural Stories	0.47***	0.27**	0.03	0.02	0.57***	0.29**	-0.01	0.00
Provo	-0.08***	0.01	-0.03	0.07	-0.07	-0.01	-0.03	-0.10**

Table 4: $\Delta_{\mathcal{L}}$ (in 10^{-2} nats) over baseline *with* surprisal when adding a similarity-adjusted surprisal or information value term for (only) the current word w_t. Monte Carlo (MC) estimation with 50 samples.