Testing the Predictions of Surprisal Theory in 11 Languages

Ethan G. Wilcox[1] **Tiago Pimentel**[2] **Clara Meister**[1] **Ryan Cotterell**[1] **Roger P. Levy**[3]
[1]ETH Zürich, Switzerland [2]University of Cambridge, UK [3]MIT, USA
ethan.wilcox@inf.ethz.ch tp472@cam.ac.uk clara.meister@inf.ethz.ch
ryan.cotterell@inf.ethz.ch rplevy@mit.edu

Abstract

Surprisal theory posits that less-predictable words should take more time to process, with word predictability quantified as surprisal, i.e., negative log probability in context. While evidence supporting the predictions of surprisal theory has been replicated widely, much of it has focused on a very narrow slice of data: native English speakers reading English texts. Indeed, no comprehensive multilingual analysis exists. We address this gap in the current literature by investigating the relationship between surprisal and reading times in eleven different languages, distributed across five language families. Deriving estimates from language models trained on monolingual and multilingual corpora, we test three predictions associated with surprisal theory: (i) whether surprisal is predictive of reading times, (ii) whether expected surprisal, i.e., contextual entropy, is predictive of reading times, and (iii) whether the linking function between surprisal and reading times is linear. We find that all three predictions are borne out crosslinguistically. By focusing on a more diverse set of languages, we argue that these results offer the most robust link to date between information theory and incremental language processing across languages.

1 Introduction

Language processing is incremental and dynamic: When a reader encounters a word, they allocate a certain amount of time to process it before moving on to the next one. One influential theory for the mechanism underlying this process is surprisal theory (Hale, 2001; Levy, 2008), which states that the time required to successfully comprehend a word is based on its predictability. Notably, predictability is often quantified as **surprisal** (negative log-probability given preceding context), from which the theory's name is derived. Suprisal theory is supported, empirically, by a number of studies which have found that surprisal is strongly correlated with psychometric measurements in large naturalistic reading corpora (Demberg and Keller, 2008; Wilcox et al., 2020; Shain, 2019, 2021; Meister et al., 2021; Pimentel et al., 2023; Hoover et al., 2022, *inter alia*). Put differently, a word's surprisal is a strong correlate of its processing effort, operationalized as reading time.

However, there is one serious limitation with most previous studies: While making general claims about human language processing, they predominantly investigate reading times in *English*. And, while a few studies have investigated surprisal effects in languages other than English, e.g., Meister et al. (2021) in Dutch and Kuribayashi et al. (2021, 2022) in Japanese, no systematic, crosslinguistic analysis has been performed. As multiple sentence processing phenomena exhibit significant crosslinguistic variation (Hillert, 1998), the extent to which surprisal theory generalizes crosslinguistically is a nontrivial limitation of the current state of the literature.

In addition, two recent contributions have posited several extensions to surprisal theory—most influentially, (a) that contextual entropy, i.e., expected surprisal, also correlates with reading times, and (b) that the relationship between surprisal and reading time is *linear* (Smith and Levy, 2013; Wilcox et al., 2020; Shain et al., 2022). Regarding (a), Pimentel et al. (2023) and Cevoli et al. (2022) have argued for what may be considered an expanded version of surprisal theory where processing difficulty is still determined by surprisal, but where people's reading behavior is additionally sensitive to expected surprisal (contextual entropy). Building off prior work that has investigated the role of entropy in language processing (Hale, 2003; Roark et al., 2009; Linzen and Jaeger, 2016; van Schijndel and Schuler, 2017), these recent studies suggest that readers may allocate reading times in advance

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1451–1470, 2023. https://doi.org/10.1162/tacl_a_00612
Action Editor: Wenjie (Maggie) Li. Submission batch: 3/2023; Revision batch: 6/2023; Published 12/2023.

of encountering a word, based on their expectations of how difficult the word will be to process. Regarding (b), a number of studies have found evidence that the linking function between reading times and surprisal is linear (Smith and Levy, 2013; Wilcox et al., 2020; Shain et al., 2022). However, these results have been challenged recently, with different studies coming to different conclusions about the most appropriate linking function. In the past two years, for example, investigations have concluded that this function is sublinear (Brothers and Kuperberg, 2021), linear (Shain et al., 2022), and superlinear (Meister et al., 2021; Hoover et al., 2022). Here, we will use the term **surprisal theory** to refer to both the core hypothesis that reading times are correlated with surprisal, as well as the two extensions—(a) and (b)—described above.

We address a gap in the current literature by investigating the predictions of surprisal theory, on eleven languages distributed across five language families.[1] We enumerate these three predictions as hypotheses below.

Hypothesis 1 ***(Surprisal Hypothesis).*** *Surprisal is predictive of reading times.*

Hypothesis 2 ***(Contextual Entropy Hypothesis).*** *Contextual entropy is predictive of reading times.*

Hypothesis 3 ***(Linear Link Hypothesis).*** *The linking function between surprisal and reading times is linear.*

We facilitate crosslinguistic comparison by using the MECO dataset (Siegelman et al., 2022), which presents eye-tracking data on reading materials with the same content in each language. We estimate surprisal and contextual entropy from two types of autoregressive language models—a single, large, multilingual model (mGPT; Shliazhko et al. 2022), as well as monolingual models trained on large and small datasets, where the small dataset is the same size across languages ($\approx$ 30 million words). We quantify the psychometric predictive power of surprisal and contextual entropy (i.e., how well each predicts reading times) by including them as variables in linear regression models. These models are then trained to predict by-word reading times; if the log-likelihood of the regression improves after including these variables, we take this as evidence that those variables have psychometric predictive power (Frank and Bod, 2011; Fossum and Levy, 2012; Goodkind and Bicknell, 2018).

We find that, in all languages tested, regression models that include surprisal are significantly better predictors of reading times over baselines which do not include surprisal, confirming the surprisal hypothesis. Additionally, we find that models which include contextual entropy are even better predictors of reading times in most languages tested, confirming the contextual entropy hypothesis. Finally, compatible with the linear link hypothesis, we find that models constrained to a linear relationship between surprisal and reading times are just as good as those that can express more complex relationships. Overall, our results provide the largest crosslinguistic analysis of the relationship between reading and word-level information theoretic properties to-date.

2 Psycholinguistic Predictive Power

Our behavior of interest is how long readers spend visually attending to a given word w_t in its linguistic context, i.e., w_t's reading time. This quantity offers a window into the psychological processes that underlie language comprehension and is typically taken as a direct reflection of the word's processing difficulty (Rayner, 1998). A word's reading time can be measured via multiple experimental modalities, including self-paced reading (Just et al., 1982; Jegerski, 2013) and the maze task (Forster et al., 2009; Boyce et al., 2020). In this work, we focus on eye-tracking measurements. These measurements have high temporal resolution and exhibit smaller spillover effects than self-paced reading (Smith and Levy, 2013), where spillover is the effect of a word's properties on later words' reading behavior.

Following previous work investigating reading, we ask what factors associated with each word are helpful for predicting its reading times. In the following section, we use the following notation. With w, we denote a word taken from an alphabet Σ. With $\boldsymbol{w} \in \Sigma^*$, we denote a string of words over the alphabet Σ. We write w_t for the word at index t in a string $\boldsymbol{w} = w_1 \cdots w_T$ with $1 \leq t \leq T$. Additionally, let $\textsc{eos} \notin \Sigma$ be a distinguished end-of-string symbol *not* in Σ and let $\overline{\Sigma} \stackrel{\text{def}}{=} \Sigma \cup \{\textsc{eos}\}$ be an augmented alphabet that

[1] Our languages (and families) are: Korean (Koreanic), Turkish (Turkic), Hebrew (Semitic), Finnish (Uralic), Dutch, English, German, Greek, Italian, Russian, and Spanish (Indo-European).

includes EOS. With each word w_t in a context $\boldsymbol{w}_{<t}$, we associate a real column vector of predictor variables $\mathbf{x}_t$ that we believe may impact reading times. Many of these predictors are attributes of w_t itself, e.g., w_t's length. We use $\mathbf{x}_t$ as predictors in a regression model $f_{\boldsymbol{\phi}}$ with parameters $\boldsymbol{\phi}$. The regression model is estimated to predict w_t's reading time from data. In symbols, we write that

$$y(w_t, \boldsymbol{w}_{<t}) \sim f_{\boldsymbol{\phi}}(\cdot \mid \mathbf{x}_t) \tag{1}$$

where $y(w_t, \boldsymbol{w}_{<t})$ is the reading time of word w_t in context $\boldsymbol{w}_{<t}$. To be explicit, in our formulation we treat reading times as a continuous quantity and, thus, $f_{\boldsymbol{\phi}}$ is a probability *density*.

In order to contrast different theories of language processing, we compare regression models with different vectors of predictor variables $\mathbf{x}$ and with different architectures $f_{\boldsymbol{\phi}}$, each of which is taken to instantiate a different hypothesis about what underlying factors determine reading times. We fit each regression model on a portion of our dataset and evaluate it by measuring the log-likelihood that it assigns to held-out data. Models that lead to higher log-likelihood can be said to have better predictive power or psychological accuracy for human reading—and their associated theories are then taken to be better models of the underlying psycholinguistic processes (Frank and Bod, 2011; Fossum and Levy, 2012).

Typically, for each experiment we will define a **target regression model**, which is trained to predict the reading times of individual words from a set of baseline predictors plus a predictor of interest (e.g., surprisal or contextual entropy). For a specific index t, we will refer to these predictors as our **target predictors** and denote them as $\mathbf{x}_t^{\text{tgt}}$. We also define a **baseline regression model** that includes only the **baseline predictors**, which are a subvector of the target predictors, denoted as $\mathbf{x}_t^{\text{base}}$ for a specific t. We denote baseline and target regression models symbolically as $f_{\boldsymbol{\phi}}(\cdot \mid \mathbf{x}_t^{\text{tgt}})$ and $f_{\boldsymbol{\phi}}(\cdot \mid \mathbf{x}_t^{\text{base}})$, respectively. Unless otherwise specified, the regression models that we use in this study are all linear. The choice to use linear linking functions, and whether this assumption is warranted, is addressed directly in Section 5. In order to assess whether the target predictors have contributed to better predictive power, we will inspect the (average) by-word difference in log-likelihood assigned by the two regression models to a held-out dataset (Goodkind and Bicknell, 2018; Wilcox et al., 2020). Following previous studies, we refer to this metric as the **delta log-likelihood** Δ, which is defined, for a specific index t, as

$$\begin{aligned}\Delta_t = &\log f_{\boldsymbol{\phi}}\left(y(w_t, \boldsymbol{w}_{<t}) \mid \mathbf{x}_t^{\text{tgt}}\right) \\ &- \log f_{\boldsymbol{\phi}}\left(y(w_t, \boldsymbol{w}_{<t}) \mid \mathbf{x}_t^{\text{base}}\right)\end{aligned} \tag{2}$$

where $y(w_t, \boldsymbol{w}_{<t})$ is the observed reading time of word w_t in context $\boldsymbol{w}_{<t}$. The complete metric Δ is the average of Δ_t over all word indices. A positive Δ means that the target predictors contribute to psycholinguistic predictive power above the baseline predictor, whereas a Δ of zero indicates that the added predictors either lack a robust relationship with reading times or that their functional relationship cannot be approximated by the class of models $f_{\boldsymbol{\phi}}$ we employ.[2] Below, we briefly introduce the two target predictors associated with the theories that we wish to test: surprisal and contextual entropy.

2.1 Surprisal

The surprisal (Shannon, 1948) of a word w_t measures the information content it conveys in the context in which it appears. Using Shannon's formulation of entropy, we can define surprisal as

$$s_t(w_t) \stackrel{\text{def}}{=} -\log_2 p(w_t \mid \boldsymbol{w}_{<t}) \tag{3}$$

where $p(\cdot \mid \boldsymbol{w}_{<t})$ is the true distribution over words $w \in \overline{\Sigma}$ in context $\boldsymbol{w}_{<t}$, which we omit from the notation for brevity. We focus here on reading, where the relevant context to compute surprisal is the w_t's preceding words $\boldsymbol{w}_{<t}$. However, in our studies, we do not have access to the true distribution $p(\cdot \mid \boldsymbol{w}_{<t})$ and instead estimate it using an autoregressive language model, as is common in previous studies (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020).

2.2 Contextual Entropy

The contextual entropy of a $\overline{\Sigma}$-valued random variable W_t at index t is the expected value of its surprisal, which can be expressed as

[2] In practice, negative values of Δ are also possible; they indicate overfitting, and imply the same theoretical conclusion as a Δ of 0.

$$\mathrm{H}(W_t \mid \boldsymbol{W}_{<t} = \boldsymbol{w}_{<t}) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{w \sim p(\cdot \mid \boldsymbol{w}_{<t})} [s_t(w)] \quad (4)$$
$$= -\sum_{w \in \overline{\Sigma}} p(w \mid \boldsymbol{w}_{<t}) \log_2 p(w \mid \boldsymbol{w}_{<t})$$

Again, as we do not have access to the true distribution p, so we resort to estimating the contextual entropy using an autoregressive language model.

Prior work has investigated the relationship between contextual entropy and reading behavior: A number of studies have investigated entropy reduction, or the extent to which w_t reduces uncertainty over possible next words (Frank, 2010, 2013) or the possible incremental parses that can be assigned to a sentence prefix (Hale, 2003, 2006). Other researchers have investigated the effect of successor entropy, i.e., the entropy of W_{t+1}, on predicting the current-word reading times (Roark et al., 2009; Linzen and Jaeger, 2016; van Schijndel and Schuler, 2017).[3] In contrast, we look at the effect of W_t's contextual entropy on prediction, following Pimentel et al. (2023) and Cevoli et al. (2022). As discussed in Pimentel et al. (2023), investigating contextual entropy separately from surprisal can uncover to what extent reading behavior is responsive (i.e., driven by surprisal) or anticipatory (i.e., driven by expected surprisal). Pimentel et al. (2023) specifically found that contextual entropy is a significant predictor of reading times on 3 out of 4 of their tested English eye-tracking and self-paced reading datasets.

3 Experimental Setup

3.1 Dataset

We use the Multilingual Eye Movement Corpus (MECO; Siegelman et al., 2022). MECO contains eye-tracking data from L1 speakers (between 29 and 54 per language) for 12 simplified Wikipedia-style articles in thirteen languages; these languages are from five different language families. Articles in the MECO corpus went through an iterative translation process by separate teams of translators to ensure that article content was the same across languages and range from a minimum 1,487 total words (Finnish) to a maximum 3,021 total words (Russian). The eleven languages we include in our analysis are: Korean (Koreanic), Turkish (Turkic), Hebrew (Semitic), Finnish (Uralic), Dutch, English, German, Greek, Italian, Russian, and Spanish (Indo-European).[4] While this sample is still biased towards Indo-European languages, it is more diverse than other previous studies, which have tended to focus exclusively on a single language.

The following pre-processing steps were taken: Words that were skipped on the first pass were given a reading-time of zero and included in the analysis. Eye-tracking datasets report multiple different word-based measurements of reading times, of which we use three (Rayner, 1998): The **first fixation** is the duration of the first fixation on a word during its first pass. **Gaze duration** is the sum of all first-pass fixations on a word. And **total fixation** time is the sum of all fixations on a word during the trial. While we report results for all three for the sake of completeness, our discussion will focus on results for gaze duration as has been done in previous studies, e.g., Wilcox et al., 2020. First fixation times are typically associated word identification (Clifton et al., 2007) and are expected to not reflect strong contextual influences. Total reading durations can be influenced by material from the right context (i.e., regressive saccades). Thus, for studies that focused on progressive movement through a text, such as ours, gaze duration is expected to be most strongly associated with first-pass processing difficulty, which is our cognitive process of interest. For each of these metrics, we fit a regression model on averages of the reading time measures taken across subjects, as has been done in previous work (Smith and Levy, 2013; Wilcox et al., 2020). This step was performed to mitigate the potentially high by-participant variance present in eye-tracking data.

3.2 Language Models

We derive surprisal and contextual entropy estimates from both monolingual and multilingual models, which we describe in greater detail below.

Monolingual Models We train monolingual transformer models using the Wiki40B dataset (Guo et al., 2020), from which we rely on the

[3] When computing W_{t+1}, it is common to treat $W_t = w_t$ as observed.

[4] The dataset also includes Norwegian and Estonian, however these are not supported by our multilingual language model and therefore excluded.

Language	Code	# Training Tokens (mil)
Dutch	du	171
English	en	1,966
Finnish	fi	89
German	ge	883
Greek	gr	57
Hebrew	he	112
Italian	it	376
Korean	ko	75
Russian	ru	488
Spanish	sp	508
Turkish	tr	48

Table 1: Training data information for our monolingual transformer models, noted as monoT(all).

training and validation splits from the original paper for each of our analyzed languages. We first fit language-specific UnigramLM tokenizers (Kudo, 2018) with a vocabulary size of 32k on the training portion of this dataset, which we then use to tokenize both the Wiki40B and MECO text into subword units. We then train two models per language, with different amounts of training data: For the **monoT(all)** variant, we train the model on the total amount of data in Wiki40B for each language; for the **monoT(30m)** variant, we subsample $\approx$ 30 million tokens from each language. For a list of the training dataset sizes for the monoT(all) models, see Table 1. We train all our models using fairseq (Ott et al., 2019), following their recommended language modeling training hyper-parameters. We use a standard decoder-only transformer with 6 layers, a context window size of 512 tokens, and shared input–output embeddings. We train our models using Adam (Kingma and Ba, 2015), with a learning rate of $5e^{-4}$, 4000 warm-up updates, and dropout of 0.1. For both of our monolingual models, as well as the multilingual model described below, per-word surprisals are computed by summing over subword unit surprisals, which is the appropriate procedure since surprisal decomposes additively over the units compromising a signal. Because of spurious ambiguity inherent in the tokenization scheme, an efficient algorithm to estimate contextual entropy over full words is unavailable to us; such an algorithm requires summing over an infinite number of sub-word combinations. Instead, we simplify this computation by estimating contextual entropy over one single step of sub-word tokens as suggested in Pimentel et al. (2023). Techniques similar to this have been employed previously in studies of entropy (Frank, 2010), e.g., to account for clitics and contractions.

Multilingual Model We use mGPT (Shliazhko et al., 2022), a multilingual autoregressive language model, which was trained with the GPT-3 architecture on 60GB of text[5] from a combination of Wikipedia and the Cleaned Common Crawl Corpus (Raffel et al., 2020).

Context Length One recent study has hypothesized that, when deriving surprisal estimates for psycholinguistic modeling, the size of the context window can bias estimates (Hoover et al., 2022). The reasoning is that short context windows could shift probability mass away from very low-frequency words, which would be better predicted from longer contexts. Therefore, we estimate surprisal and contextual entropy from mGPT in two contexts: In short contexts the model is given only the current sentence (up until the current word); in long contexts we use the model's full input window size of 512 characters. We use long contexts for our first analysis, and use both contexts for our second analysis, which investigates both the shape of the reading times–surprisal linking function and the influence of context length on these results.

Psychological Plausibility Increasingly, researchers that use language models for cognitive modeling have considered their psychological plausibility as estimates of humans' internal notions of word predictability. In particular, some researchers have compared the size of the models' training data to the amount of linguistic experience of the average human child (Zhang et al., 2021). Assuming that children are typically exposed to $\approx$ 11 million words per year as an upper limit (Hart and Risley, 1995), then the mGPT model is trained on multiple human lifetimes' worth of language data. The monoT(all) models are trained on data scales equivalent to or less

[5]Shliazhko et al. (2022) report that their combined dataset contains 489 billion characters. Assuming a crosslinguistic average of $\approx$ 5 characters per word, this puts their training set at slightly under 100 billion words.

than one human lifetime,[6] and the monoT(30m) models are trained on data equivalent to the linguistic exposure of a young child. However, we argue that the psychological plausibility of a model's next-word predictions is not completely determined by whether that model's training data is the same size as the amount of data a human learner is exposed to. Indeed, there is a body of evidence suggesting that, beyond a certain minimal amount of data, the more data a model is trained on, the more human-like that model's next-word predictions become (Goodkind and Bicknell, 2018; Wilcox et al., 2020). All of our models are trained on an amount of data within this range. However, at the other end of the scale, the relationship flips: Models trained on an extremely large amount of data seem to be slightly *worse* predictors of human reading (Shain et al., 2022; Oh and Schuler, 2023). For our models, training datasets are uni-modal (i.e., language only) and learning is with arguably weaker priors for language-like structure, whereas humans learn from multi-modal data with potentially much stronger priors for linguistic structures. Likely, more data makes up for the lack of multi-modal data and uninformative priors.

3.3 Regression Models

All of our regression models are fit to predict the reading time $y(w_t, \boldsymbol{w}_{<t})$ of a word w_t in a context $\boldsymbol{w}_{<t}$ from the predictor vector $\mathbf{x}_t$. In addition to looking at the word w_t, our predictor includes quantities derived from the previous two words w_{t-1}, w_{t-2} to control for potential spillover effects. We will refer to the three words w_t, w_{t-1}, w_{t-2} as our **regressor words**. Following previous work in this area, all regression models include the word length and log-unigram frequency, as estimated by Speer (2022), for all regressor words in a predictor $\mathbf{x}_t$ for a specific index t. The predictors above constitute our (context invariant) baseline predictors. Regression models are trained and evaluated using 10-fold cross validation. For more information on the regressions used in each of our experiments, see Appendix A. The significance of the observed Δ values between target and baseline models is assessed via a paired permutation test that checks whether Δ is significantly different from zero. We use permutation tests for our comparisons because they make no assumption about the distribution of the test statistic. Instead, the test uses the empirical distribution of differences in likelihoods, as estimated using averages computed over permutations of likelihoods, in order to compute p-values.

[6]The only exception is English, which at $\approx$ 2 billion words is about two lifetime's worth of linguistic data, assuming the 11-million word per year estimate of Hart and Risley (1995).

4 Results

4.1 Surprisal

To test the surprisal hypothesis, we fit a target regression model whose predictors includes the surprisals of our regressor words plus our baseline predictors described above. We compare this to a baseline that does not include the surprisal predictors. For this and subsequent tests, we calculate results for each language individually, as well as for the combined data from all languages. Results can be seen in Figure 1 broken down by language, model, and each of our three word-based measurements of reading time. We observe a clear pattern in the results across the languages: Positive Δ in nearly every test for gaze duration and total fixation, and less consistently positive Δ for first fixation, where, as noted before, we would not necessarily expect surprisal effects to show up. Looking at the results for each model, we observe the most robust results for mGPT, where Δ is significantly greater than zero in every language for gaze duration and total fixation. For the monolingual models, we observe more robust effects for the monoT(all) model over the monoT(30m) model, which is sensible given the latter's limited training data size.

For an aggregate test of the effects of surprisal, we fit an additional regression model on the combined data from all languages to predict gaze duration with random by-language effects. We use a fully maximal random effect structure, as advocated in Barr et al. (2013). We find that the model with surprisal leads to significantly greater than zero Δ in all cases ($p < 0.001$). Although surprisal leads to a positive Δ across languages, we do observe some variation in the magnitude of this effect, or the predictive power obtained by regression model. For both mGPT and monoT(all) we observe the highest predictive power in Russian and Dutch, with lower predictive power in Spanish, English, and Hebrew.

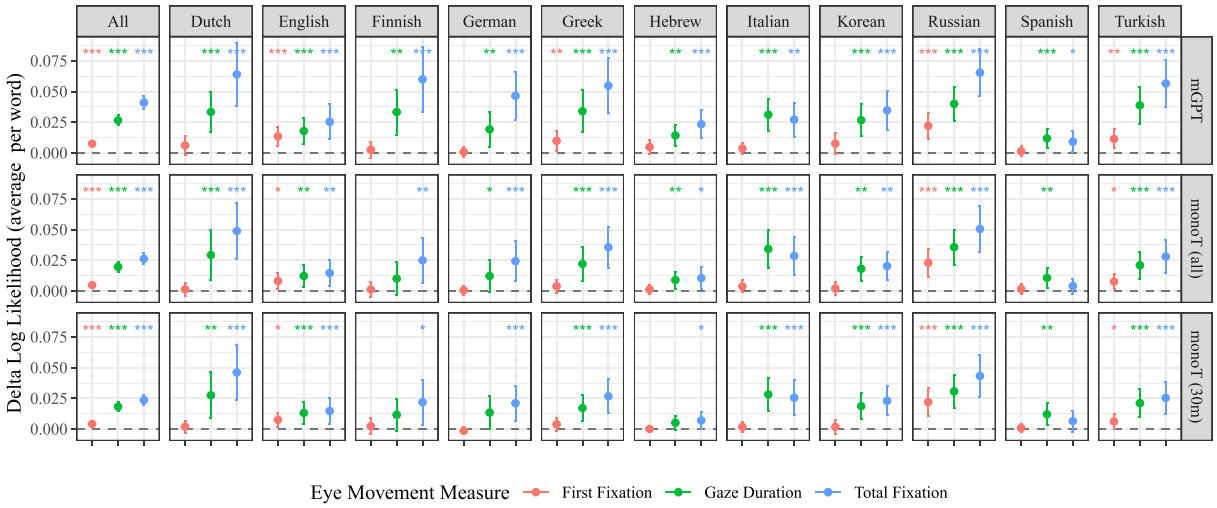

Figure 1: **Predictive Power of Surprisal Across Languages:** Positive values mean surprisal contributes to predicting the reading times over a baseline where surprisal is removed. Error bars indicate 95% confidence intervals. Asterisks indicate the significance of a paired permutation test. We find a consistent significant effect of surprisal across languages for language models that are both multilingual (top row) and monolingual (bottom two rows), and for both progressive gaze duration and total fixation.

One natural question to ask is whether imbalances in the model's training data leads to some of this variation—do models make better predictions for language where they have seen more data? However, there are converging pieces of evidence from our data suggesting that differences in dataset size is *not* the main cause of the by-language variation. First, both mGPT and monoT(all) show relatively lower predictive power for some large-data languages such as Spanish and English. Second, and quite interestingly, similar patterns of predictive power can be observed for our monoT(30m) models, where training dataset size is controlled across languages. Here, as with the other models, we observe larger values of Δ in Dutch and Russian and smaller values of Δ in English, Spanish, and Hebrew. These results pose a puzzle, as the languages for which the models obtain higher Δ are not obviously different from those for which the models obtain lower Δ, in terms of their linguistic features. For example, English (lower Δ) and Dutch (higher Δ) are both Western Germanic. Further investigation is needed to determine if these patterns hold up for other crosslinguistic reading time datasets.

4.2 Contextual Entropy

To test the contextual entropy hypothesis we first fit a single baseline regression model. Our baseline regression model includes the surprisal of all regressor words, plus baseline predictors. We then evaluate target regression models in two variants: For the *replace* regression model, we replace surprisal with contextual entropy for all regressor words. For the *add* regression model, we add an additional term of contextual entropy for all regressor words. As results do not change much between our monolingual language models, we present results for monoT(all).

Results can be seen in Figure 2, where the replace regression is indicated with a triangle and the add regression is indicated with a circle. First, we find that replacing surprisal with entropy tends to hurt predictive power in most cases. For example, for mGPT and gaze duration, Δ is negative in 6/11 languages and significantly so in two, Dutch ($p < 0.05$) and Italian ($p < 0.05$), implying overfitting. Negative effects are even stronger for the monoT(all) model, where we find negative gaze duration Δ in every language (results are significant in 5/11). Adding entropy as an *additional* predictor, on the other hand, generally improves the model's predictive power. For example, for mGPT and gaze duration, Δ from the add regression is positive in 8/11 languages, and significantly so in 5 (English, Greek, Korean, Russian, and Turkish). In addition, Δ is significantly positive for the add regression for all three reading time measures when data is combined across languages, as shown in the 'All'

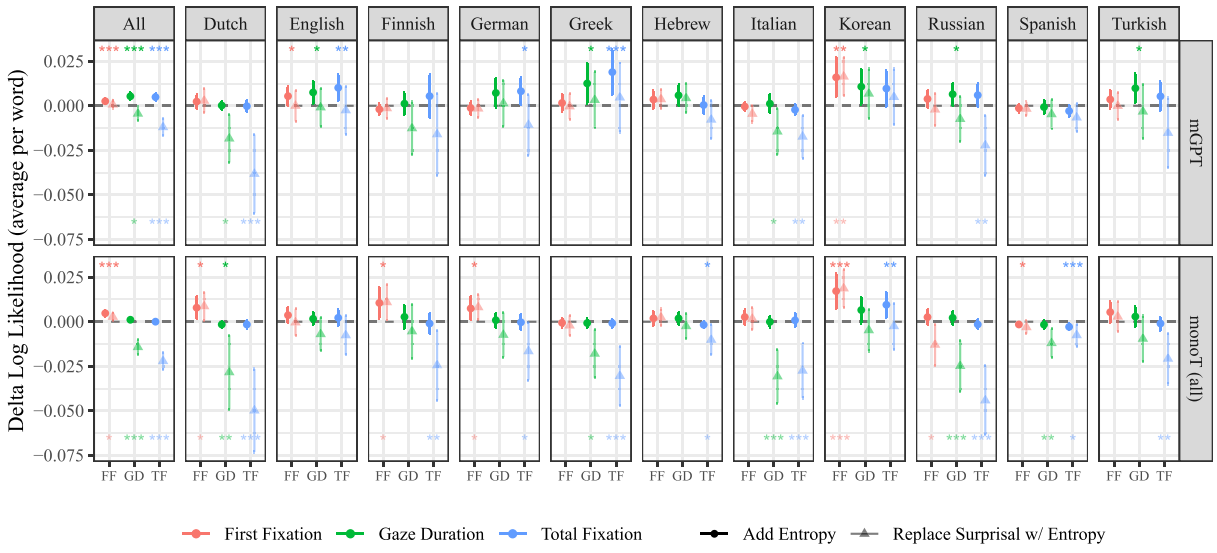

Figure 2: **Psychometric Predictive Power of Contextual Entropy Across Languages:** Positive values mean contextual entropy contributes to predicting the reading times of w_t. Error bars are 95% confidence intervals across the 10 folds of held-out data. Asterisks indicate the significance of a paired permutation test. We find that replacing surprisal with entropy tends to hurt predictive power, while adding entropy tends to help.

column at the left of Figure 2. Results are less strong for monoT(all), where positive Δ shows up predominantly for first fixation. As before, we run an aggregate test with data from all languages including by-language random effects.[7] For gaze duration, we find that adding contextual entropy leads to positive Δ (mGPT, $p < 0.001$; monoT(all), $p < 0.01$) and that replacement leads to negative Δ (mGPT, $p < 0.01$; monoT(all), $p < 0.001$). Overall, we take these results as being in line with those reported in Pimentel et al. (2023). Our findings suggest that contextual entropy has a weak—albeit consistent—effect on reading times across languages, and therefore that participants may be pre-planning their processing times based on the expected surprisal of upcoming words.

4.3 Variation Across Languages

The crosslinguistic relationship between Δ and language model quality is relevant to current debates about about whether language models can plausibly be used to understand psycholinguistic processes. As mentioned in Section 3.2, it has been observed that, within English, models with lower perplexity tend to exhibit better predictive power (Goodkind and Bicknell, 2018; Wilcox et al., 2020). However, studies on Japanese have failed to replicate these results, suggesting that the relationship does not hold for all languages (Kuribayashi et al., 2021). Further, Oh and Schuler (2023) and Shain et al. (2022) show that this relationship may not hold even in English for the most recent language models. To investigate this, we compute, for mGPT, the Pearson's correlation between Δ and test set perplexity, as reported in Shliazhko et al. 2022, both across languages, as well as across language families.[8] For this analysis we show results only for mGPT and leave a full analysis, comparing different monolingual models, for future work.

The correlations can be seen in Figure 4. We do find a relatively strong negative correlation across languages, however it is not significant ($\rho = -0.497, p = 0.1$). We do not find any evidence of correlation in the language family data. Although the negative by-language correlation suggests that, for languages where mGPT has lower perplexity, it may be a better model of psycholinguistic behavior, the lack of significance is in line with the negative results from Japanese.

[7] Following the same methodology as the previous test, we look at the effect of adding or replacing surprisal across all regressor words.

[8] For language families, Δ and perplexities are within-family averages.

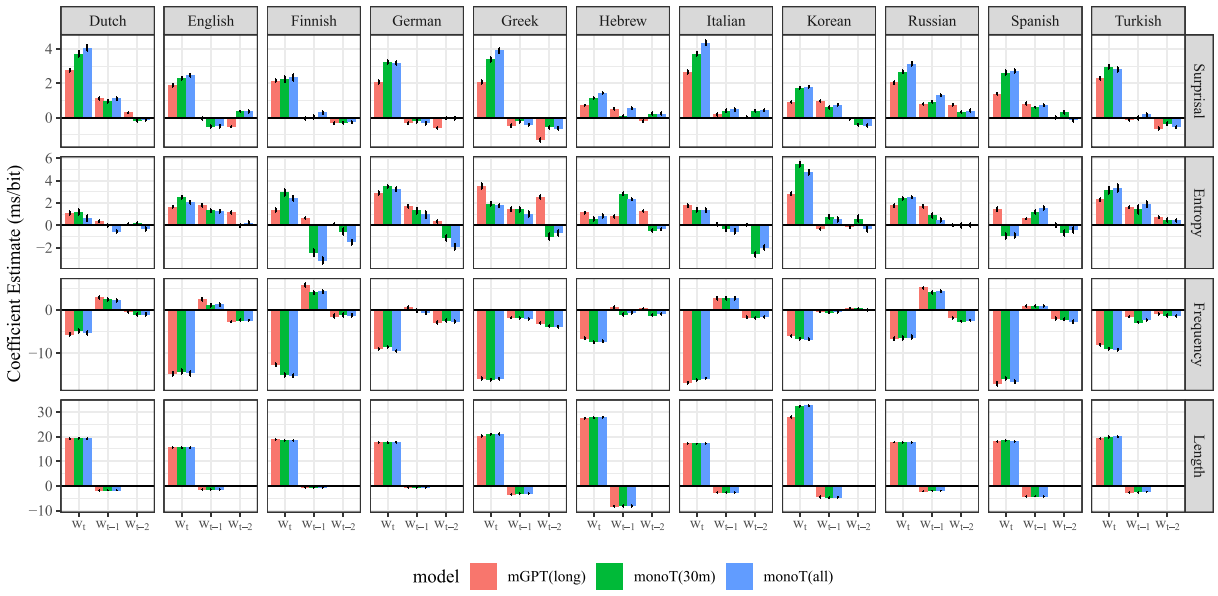

Figure 3: **Model Coefficients:** Coefficients for a linear model that includes surprisal, entropy, frequency, and length. Coefficients are shown for each regressor word individually. Zero is indicated with a black line and scales differ for each row. Error bars indicate 95% CIs across folds of data.

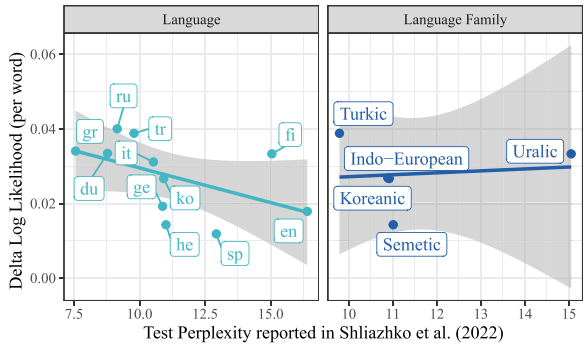

Figure 4: **Test Perplexity versus Δ (mGPT):** We do not find a significant correlation between the Δ and mGPT's perplexity for a language or language family.

Notably, there are important differences between this analysis and the studies cited above, which train a number of different language models within a single language and a single shared vocabulary, as opposed to comparing the outputs of a single multilingual language model across languages as we do here. Additionally, although mGPT does share a single vocabulary across languages, different languages might be *a priori* harder or easier to language-model (Cotterell et al., 2018; Mielke et al., 2019), and quality of the tokenization might vary across languages as well. Thus, more fine-grained linguistic controls are necessary before making strong conclusions about the relationship between perplexity and psychometric predictive power across languages.

4.4 Model Coefficients

How do surprisal, entropy, frequency, and length individually affect reading times? Figure 3 shows the estimates from regression models for each of our predictor variables, estimated across 10 folds of data. Unlike the figures presented above, effects are broken down by the coefficients for each of our regressor words from w_t (on the left of each facet) to w_{t-2} (on the right of each facet). Note that effect size here does not correspond to the predictive power of the model as a whole, but rather the impact of word-level properties on reading times. Because predictor variables are not normalized, units are different across rows. The top two rows indicates the estimated slowdown in milliseconds for each additional bit (of surprisal or entropy). The second row indicates slowdown for each additional occurrence per billion words of text (on a log scale). And the bottom row indicates slowdown for each additional character in the word.

We find a consistent effect of surprisal for w_t of between 2-4 ms/bit. There is some inter-language variability, with the smallest effect for Hebrew, and larger effects for Dutch, Russian, Greek, and Italian. We find smaller effects for

w_{t-1}, ranging from between 0-2 ms/bit. There is no obvious effect of surprisal for w_{t-2}. Overall, these results differ slightly from those reported in Smith and Levy (2013), who investigate reading times on the English Dundee Corpus (Kennedy et al., 2003) and find a stronger effect for w_{t-1} than we do. However, our results are not inconsistent with the relatively lower spillover effects traditionally observed in eye-tracking data.

Turning to contextual entropy, we find slightly smaller effects, and slightly more variance, between languages. There is no obvious relationship between the effect sizes for surprisal and contextual entropy. For example, Dutch, which has a larger surprisal effect, has one of the smallest effect sizes for entropy. For frequency, we find a consistently negative effect for w_t, as expected—as words get more frequent they take less time to read. For w_{t-1} and w_{t-2} effects are much smaller and less consistent across languages. For example, Dutch, Finnish, Italian, and Russian all have consistently positive frequency effects for w_{t-1}, whereas in Turkish and Greek, these effects are negative.

We find consistent effects for word length, which are positive for every language on w_t. We also find consistent negative effects for w_{t-1}. This may be due to the fact that readers are likely to skip a word if it comes after a long word, which would be associated with a reading time of zero in our analysis. Overall, these coefficient estimates are in line with previous reading time studies and further highlight the crosslinguistic consistency of our results.

5 Surprisal–RT Linking Function

The regression models we have been using to assess Δ have implicitly assumed a linear linking function between surprisal and reading time—a relationship that has been empirically verified in some previous studies in English (Smith and Levy, 2013; Wilcox et al., 2020; Shain et al., 2022). Other recent studies, however, have questioned linearity, including Meister et al. (2021) and Hoover et al. (2022), who argue for a *super*-linear relationship, and Brothers and Kuperberg (2021), who argue for a *sub*linear relationship. In this section, we directly test the linear link hypothesis. We compare the Δ of our linear regression models against regression models that can capture non-linear relationships. We present results exclusively for gaze duration for the reasons discussed in Section 3.1.

5.1 Visualizing the Link with GAMs

In order to visualize the link between surprisal and reading times, we use generalized additive models (GAMs), a class of models that can fit non-linear relationships between predictor and response variables. Given the less-constrained hypothesis space of the GAM, if the model finds a relationship that is (visually) linear, this is good first evidence that the underlying effect is linear. We fit a GAM to predict reading times from word frequency, length and surprisal, derived for short contexts (sentence level) and long contexts (document level). We include smooth terms for current and previous word surprisal, as well as tensor product terms for a non-linear interaction between log-frequency and word length. By way of comparison, we also fit a GAM that enforces a linear effect of surprisal, following Hoover et al. (2022). For this comparison, we fit new models, all using the `mgcv` library, as opposed to simply comparing GAMs to our linear models from the previous section, to ensure that the effects of our baseline variables are exactly the same between models in this section.[9] For each language and language model combination, we visualize the fitted curve using 10-fold cross validation, i.e., we train a GAM model on 9 of the 10 folds and sample reading times from the trained model. To sample reading times, we vary the surprisal values for w_t ranging 0–20 in increments of 0.1. No other predictors are fed into the model.

The visualizations of the estimated GAMs for effects on w_t can be seen in Figure 5. Below the fit, we show density plots for surprisal values in the corpus. The results are consistent across languages and contexts. Visually, the non-linear GAMs capture the effect of surprisal on reading times by fitting an approximately linear curve, which sometimes falls directly on top of the linear control GAM (e.g., for Finnish and Turkish). Unlike Hoover et al. (2022) we do not find a consistent difference for fits between surprisals derived in short contexts versus long contexts.

[9] For these analyses we choose to only include surprisal, frequency, and length from w_t and w_{t-1} as predictors. This was done because of the minimal effects found on w_{t-2} in our analysis of coefficients (see Figure 3). A sample GAM call for this analysis is given in Appendix A.

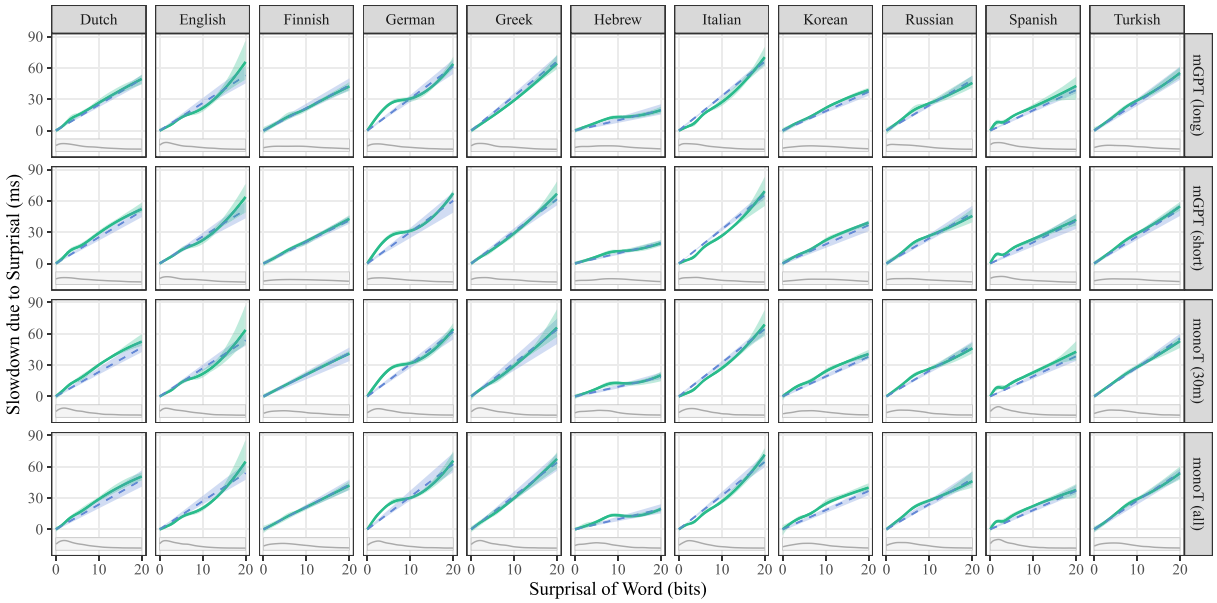

Figure 5: **Surprisal versus Reading Time Relationship:** Non-linear GAMs are in green and linear control GAMs are in dotted blue. Shaded regions represent bootstrapped 95% confidence intervals. Results are for gaze duration. Grey subplots indicate the distribution of surprisal values. We find that GAMs recover a linear relationship between surprisal and reading-time slowdown.

We note, however, that Hoover et al. (2022) find superlinear trends specially for their best examined models (e.g., GPT-3), which may outperform multilingual mGPT.

5.2 Testing Linearity

Although the GAM fits in Figure 5 are *visually* linear, we would like to test the question of linearity with a more rigorous method. To do so, we compare the Δ of the linear and non-linear GAMs described above. Δ is calculated by comparing each model to a shared baseline that includes only tensor product terms for frequency and length. The idea is that if the underlying relationship between surprisal and reading time is non-linear, then the non-linear GAMs should be able to achieve higher Δ, whereas if the underlying relationship is linear then the non-linear GAMs would not have an advantage. Thus, a consistently null result across languages suggests that the relationship is linear.

The results of this comparison can be seen in Figure 6. Here, Δ is slightly different for linear models than in Section 4.1, as we fit these models with tensor product terms for baseline predictors. Visually, there is no consistent difference between linear and non-linear models across languages. We test the difference in Δ statistically with permutation tests, as described in Section 3.3. Our tests do not support the alternative hypothesis for an $\alpha = 0.05$ for any of the models or languages. Together with the visualizations presented above, these results support a linear linking function between surprisal and reading times.

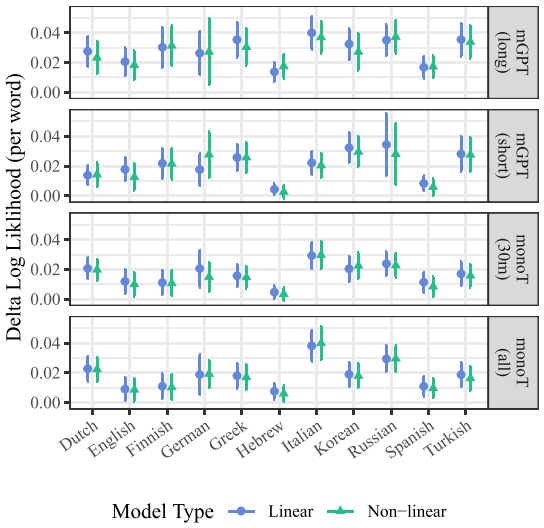

Figure 6: **Comparison Between Linear and Non-linear Models:** Error bars are 95% CIs of Δ. Results are for gaze duration. We observe no difference between non-linear GAMs (green) and linear GAMs (blue) across languages.

6 Discussion

6.1 Implications of Psycholinguistic Theories

Throughout the paper, we have mentioned that the eleven languages studied come from five different language families, but what does this mean in terms of the actual linguistic characteristics that they exhibit? At the highest organizational level, our sample includes languages with multiple different word orders and headedness including SVO (Hebrew, English), SOV (Korean, Turkish), as well as languages with no dominant word order (German and Greek; Haspelmath et al., 2005). Our sample includes languages with extensive case marking such as Finnish (15 cases), as well as languages with extremely impoverished case systems, such as English. In terms of word construction, our sample includes languages that are both agglutinating (Turkish, Finnish and Korean) and fusional (Russian, Romance languages). While this set is not close to covering all ways that human languages can vary, we bring up these differences to highlight how it does contain important high-level parametric variations observed in human languages.

In light of this, the stability observed in our results testing the surprisal hypothesis is rather remarkable. Across language families and model types, we observe essentially consistent results, in terms of the predictive power of the models, the effect size associated with surprisal, as well as for the shape of the surprisal–reading-time relationship. Focusing first on predictive power, we find a relatively tight range of Δ values associated with surprisal. For example, for gaze duration and mGPT, all Δ values fall between 0.012 and 0.040. Indeed, across languages and models, we find relatively little variance in the predictive power of surprisal. Turning to the effect size of surprisal, we observe a millisecond-per-bit tradeoff that falls between 2–4 ms/bit for every language (see Figure 3). The previous estimate of 3.75 ms of slowdown per bit of surprisal reported in Smith and Levy (2013) for English falls well within this range (though note that this previous work used surprisal estimates derived from an n-gram model, which will generally be higher than surprisal estimates derived from large neural language models such as the ones we consider in this study). We take these results to suggest that humans may have stable crosslinguistic preferences for the rate at which they process information during reading, i.e., not greater than 4 milliseconds per bit of information. This is consistent with previous work that has observed cross-linguistic consistency in the rate of information during speech production (Pellegrino et al., 2011; Coupé et al., 2019), as well as trade-offs between the information content of a word and the time taken to produce it (Pimentel et al., 2021).[10]

One point of difference between these and previous results, however, is the size of the effect of the surprisal of previous words. Looking at gaze duration in the Dundee corpus of English (Kennedy et al., 2003), Smith and Levy (2013) find an effect on reading time for surprisal for the previous word which is about as strong as for the current word. We find much weaker effects in this study, ranging from 0-2 ms/bit. Note, that this lower effect for previous words is in line with other incremental processing measures which are strongly incremental, such as the maze task, where previous-word surprisal has little to no effect on reading time of the current word (Boyce and Levy, 2020), as well as with the results reported in Pimentel et al. (2023) for eye-tracking over the Provo (Luke and Christianson, 2018) and Dundee corpora.

Turning to the shape of the surprisal–reading times relationship, our results support the linear link hypothesis and are in line with the comprehensive results recently reported in Shain et al. (2022). Unlike Hoover et al. (2022) we do not observe superlinear surprisal–reading time relationships for larger and more data-intensive language models, or for language models that had access to longer contextual windows. Interestingly, we do observe that the one language which visually appears to be superlinear, (i.e., it has an upwards curve in Figure 5) is English. Thus, while we believe Hoover et al. (2022) was right to be concerned by a potential visual nonlinearity in the English relationship, this effect does not appear to exist crosslinguistically and is not borne out by our statistical testing.

[10] Our results are not necessarily consistent with a universal channel *capacity*, or an information rate above which comprehension cannot be sustained. A channel capacity could explain uniform information density effects, or the tendency to spread information out uniformly over a sentence, presumably at or near the channel capacity (Levy and Jaeger, 2006; Frank and Jaeger, 2008; Meister et al., 2021). However, as pointed out in Smith and Levy (2013), such effects require a superlinear surprisal link hypothesis, which we do not observe empirically.

Surprisal theory is attractive because it offers a general-purpose link between statistical properties of natural language and human behavior. While its domain generality gives the theory a universal-like flavor, previous literature has (in our opinion) correctly refrained from overtly discussing it as a universal of human language processing. By conducing the most comprehensive crosslinguistic assessment of surprisal theory to date, this study presents initial evidence which supports the universality of surprisal effects in naturalistic reading. That being said, further testing is a necessary next step.

6.2 Implications of Multilingual Language Modeling

As the number of multilingual language models has proliferated, it has become increasingly important to understand how they differ from more traditional, monolingual models. Previous studies have produced mixed results: Some have found that the larger training data scales of multilingual models leads to better performance (Conneau et al., 2020), while others have found advantages for monolingual models (Agerri et al., 2020; Rönnqvist et al., 2019; Virtanen et al., 2019), which are often attributed to monolingual models' language-specific tokenization and vocabulary representation. The majority of these previous studies have focused on masked language models (mostly using architecture based off the BERT model) and evaluation based on performance of downstream tasks (Doddapaneni et al., 2021). This study offers a useful complement to previous work by focusing on autoregressive models, as well as on their cognitive modeling capacities.[11] Our results are more or less in line with previous studies, insofar as we find no obvious differences between our multilingual model and our monolingual models. Our results thus suggest that for computational linguists interested in cognitive modeling, multilingual and monolingual language models may be equally viable options. However, we would like to note that we did not compare models in truly low-resource settings, as the training datasets of our smallest monolingual models still included 30 million tokens. It may be the case that when trained on much smaller datasets, multilingual models may benefit from crosslingual transfer.

6.3 Concurrent Work

We want to briefly note the differences between the work presented here and a concurrent study that also used the MECO dataset (i.e., de Varda and Marelli, 2022). While de Varda and Marelli's research questions are similar to ours, their methods and conclusions are quite different. Instead of an autoregressive language model, they use a masked language model (mBERT; Devlin et al., 2019), which has access to both left and right context. An issue with this strategy is that the surprisal values produced by this setup are not psychologically plausible estimates of actual surprisals, which are estimated from the left context alone.[12] which weakens the ability to test psycholinguistic causal claim about the relationship between surprisal and reading times. In their experiments, de Varda and Marelli do not find significant effects of pseudo-surprisal on gaze duration in four of the 12 languages in MECO,[13] including English, and find significant effects of pseudo-surprisal on other eye movement measures in even fewer of the languages, which they view as evidence that surprisal might *not* be a consistent predictor of reading times across languages.[14] While we are aligned on the importance of de Varda and Marelli's research questions, we believe that their failure to replicate surprisal effects for English—or to find it for other languages—reflects the limitations in their methodological choices.

6.4 Limitations and Future Directions

Turning back to our own study, there are a few limitations we would like to discuss: Although our sample of languages is much larger than previous studies, Indo-European languages are still overrepresented. Indeed, each of our non Indo-European language families is represented by a

[11] However, see Hollenstein et al. (2021) for a previous investigation of multilingual language models' ability to predict reading times.

[12] Because the perceptual span is limited to about 14 characters to the right of a fixation (Rayner, 1975) and little linguistic information is gleaned from the far right of the perceptual span (Schotter et al., 2012), upcoming word identities cannot have a substantial causal influence on a word's first-pass reading behavior (Granger, 1969).

[13] They include Estonian, which we drop as it was not in mGPT's training data.

[14] Their study does not consider contextual entropy.

single language. Additionally, all the data tested here comes from high-resource languages with long traditions of writing systems, and from individuals who live in industrialized societies. Finally, the methodology we employ here requires a large corpus of (written) language on which a language model can be trained. It may be the case, that for much lower-resource languages, there is often not enough linguistic data to derive statistical estimates needed to test surprisal theory in this manner. Thus, while our methods may be able to test the predictions of surprisal theory in lower-resource settings, where corpora of a few hundred thousand words exist, they may not be suitable for a large number of the world's languages. While our results put surprisal theory on firmer empirical footing, testing its predictions beyond these settings is an important and necessary step in assessing the theory's universality.

7 Conclusion

This paper has presented the most comprehensive crosslinguistic evaluation of surprisal theory reported in the literature to date. Using eye-tracking data from controlled materials in eleven languages across five language families, we have tested three hypotheses: (i) the surprisal hypothesis (surprisal is predictive of reading times), (ii) the contextual entropy hypothesis (contextual entropy is predictive of reading times), and (iii) the linear link hypothesis (the relationship between surprisal and reading times is linear). We found exceptionally strong crosslinguistic stability in our results, with each prediction being borne out in every language tested. These results provide the most robust link between information-theoretic quantities and incremental processing.

Acknowledgments

We would like to thank our TACL action editor, Maggie Li, as well as our reviewers, whose thoughtful feedback greatly improved this work. T.P. was supported by a Facebook PhD Fellowship. C.M. was supported by the Google PhD Fellowship. E.G.W. was supported by an ETH Zurich Postdoctoral Fellowship. R.P.L. was supported by NSF grant BCS-2121074 and a Newton Brain Science Award.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: The case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278. https://doi.org/10.1016/j.jml.2012.11.001, PubMed: 24403724

Veronica Boyce, Richard Futrell, and Roger P. Levy. 2020. Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082. https://doi.org/10.1016/j.jml.2019.104082

Veronica Boyce and Roger Levy. 2020. A-maze of natural stories: Texts are comprehensible using the maze task. In *Talk at 26th Architectures and Mechanisms for Language Processing conference (AMLaP 26)*. Potsdam, Germany.

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. https://doi.org/10.1016/j.jml.2020.104174, PubMed: 33100508

Benedetta Cevoli, Chris Watkins, and Kathleen Rastle. 2022. Prediction as a basis for skilled reading: Insights from modern language models. *Royal Society Open Science*, 9(6):211837. https://doi.org/10.1098/rsos.211837, PubMed: 35719885

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. *Eye Movements*, pages 341–371. https://doi.org/10.1016/B978-008044980-7/50017-3

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle

Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2085

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):1–10. https://doi.org/10.1126/sciadv.aaw2594, PubMed: 32047854

Andrea de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144, Online only. Association for Computational Linguistics.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210. https://doi.org/10.1016/j.cognition.2008.07.008, PubMed: 18930455

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*.

Kenneth I. Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, 41(1):163–171. https://doi.org/10.3758/BRM.41.1.163, PubMed: 19182136

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 61–69, Montréal, Canada. Association for Computational Linguistics.

Austin F. Frank and T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Stefan Frank. 2010. Uncertainty reduction as a measure of cognitive processing effort. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 81–89, Uppsala, Sweden. Association for Computational Linguistics.

Stefan L. Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5(3):475–494. https://doi.org/10.1111/tops.12025, PubMed: 23681508

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834. https://doi.org/10.1177/0956797611409589, PubMed: 21586764

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. https://doi.org/10.18653/v1/W18-0102

C. W. J. Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438. https://doi.org/10.2307/1912791

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

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*. https://doi.org/10.3115/1073336.1073357

John Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32:101–123. https://doi.org/10.1023/A:1022492123056, PubMed: 12690827

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4). https://doi.org/10.1207/s15516709cog0000_64, PubMed: 21702829

Betty Hart and Todd R. Risley. 1995. *Meaningful Differences in the Everyday Experience of Young American Children*. Paul H. Brookes Publishing Co, Baltimore, MD.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. 2005. *The World Atlas of Language Structures*. Oxford University Press.

Dieter Hillert. 1998. *Sentence Processing: A Crosslinguistic Perspective*. Brill. https://doi.org/10.1163/9780585492230

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.10

Jacob Louis Hoover, Morgan Sonderegger, Steven T. Piantadosi, and Timothy J. O'Donnell. 2022. The plausibility of sampling as an algorithmic theory of sentence processing. *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/qjnpv

Jill Jegerski. 2013. Self-paced reading. In *Research Methods in Second Language Psycholinguistics*, pages 36–65. Routledge. https://doi.org/10.4324/9780203123430

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. https://doi.org/10.1037/0096-3445.111.2.228

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1007

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.712

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. 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 5203–5217, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.405

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006, PubMed: 17662975

Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, 19. https://doi.org/10.7551/mitpress/7503.003.0111

Tal Linzen and T. Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 40(6):1382–1411. https://doi.org/10.1111/cogs.12274, PubMed: 26286681

Steven G. Luke and Kiel Christianson. 2018. The Provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50:826–833. https://doi.org/10.3758/s13428-017-0908-4, PubMed: 28523601

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*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.74

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1491

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. https://doi.org/10.1162/tacl_a_00548

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4009

François Pellegrino, Christophe Coupé, and Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language*, 87(3):539–558. https://doi.org/10.1353/lan.2011.0057

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.73

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):1–67.

Keith Rayner. 1975. The perceptual span and peripheral cues in reading. *Cognitive Psychology*, 7(1):65–81. https://doi.org/10.1016/0010-0285(75)90005-5

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372. https://doi.org/10.1037/0033-2909.124.3.372, PubMed: 9849112

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. https://doi.org/10.3115/1699510.1699553

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Elizabeth R. Schotter, Bernhard Angele, and Keith Rayner. 2012. Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, 74(1):5–35. https://doi.org/10.3758/s13414-011-0219-2, PubMed: 22042596

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. https://doi.org/10.18653/v1/N19-1413

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. https://doi.org/10.18653/v1/2021.acl-long.288

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time. *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/4hyna

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mGPT: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, Marco Marelli, Timothy C. Papadopoulos, Athanassios Protopapas, Satu Savo, Diego E. Shalom, Natalia Slioussar, Roni Stein, Longjiao Sui, Analí Taboh, Veronica Tønnesen, Kerem Alp Usal, and Victor Kuperman. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863. https://doi.org/10.3758/s13428-021-01772-6, PubMed: 35112286

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319. https://doi.org/10.1016/j.cognition.2013.02.013, PubMed: 23747651

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Marten van Schijndel and William Schuler. 2017. Approximations of predictive entropy correlate with reading times. In *Proceedings of the Cognitive Science Society*, pages 1260–1265.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

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

In *Proceedings of the 2020 Meeting of the Cognitive Science Society*, pages 1707–1713, Online. Cognitive Science Society.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? 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 1112–1125, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.90

A Regression Modeling Details

We give more details on the regression formulae used in the various experiments reported in the main section. Our notation is as follows: `reading_time` is the reading time of the word of interest, i.e., w_t, `surp` is the surprisal of w_t, `prev_surp` is the surprisal of the previous word, i.e., w_{t-1}, and `prev2_surp` is the surprisal of the word two previous, i.e., w_{t-2}. The other variables use the same `prev` and `prev2` prefixes, and we simply explain the variable names for the current index t for the sake of brevity, below. For these, `ent` indicates the contextual entropy of W_t, `len` indicates the length of w_t in characters, and `freq` indicates the log unigram frequency of w_t.

Effect of Surprisal (Section 4.1) For the tests assessing the effect of surprisal within individual languages, we use the following model:

```
lmer(reading_time ~ surp + prev_surp + prev2_surp + freq + len + prev_freq +
    prev_len + prev2_freq + prev2_len, data = .)
```

The baseline models are the same with the exception that the surprisal terms are removed. For the aggregate test assessing the effect of surprisal across languages, we use the following model:

```
lmer(reading_time ~ surp + prev_surp + prev2_surp + freq + len + prev_freq +
    prev_len + prev2_freq + prev2_len + (surp + prev_surp + prev2_surp + freq
    + len + prev_freq + prev_len + prev2_freq + prev2_len | lang), data = .)
```

Effect of Contextual Entropy (Section 4.2) For both tests, the baseline model included surprisal, length and unigram frequency, i.e., it was the first model given in the paragraph above. For the *replace* test, the target regression model we use is

```
lmer(reading_time ~ ent + prev_ent + prev2_ent + freq + len + prev_freq +
    prev_len + prev2_freq + prev2_len, data = .)
```

For the *add* test, the target regression model we use is

```
lmer(reading_time ~ ent + prev_ent + prev2_ent + surp + prev_surp +
    prev2_surp + freq + len + prev_freq + prev_len + prev2_freq + prev2_len,
    data = .)
```

Surprisal–RT Linking Function (Section 5) The GAM formula used for non-linear models we use is

```
gam(reading_time ~ s(surp, bs = 'cr', k = 6) + s(prev_surp, bs = 'cr',
    k = 6) + te(freq, len, bs = 'cr') + te(prev_freq, prev_len, bs = 'cr'),
    data = .)
```

And for linear models:

```
gam(reading_time ~ surp + prev_surp + te(freq, len, bs = 'cr') + te(prev_freq,
    prev_len, bs = 'cr'), data = .)
```

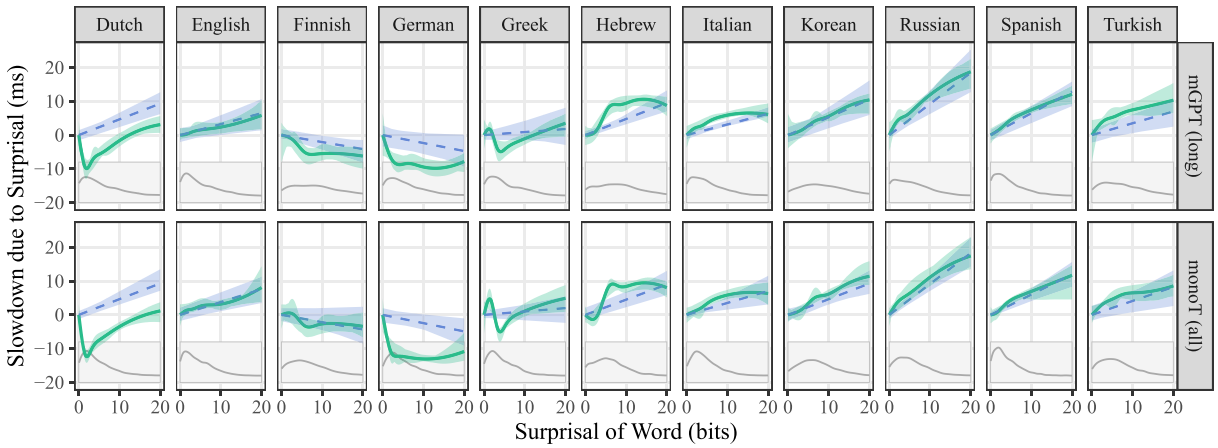

Figure 7: **Surprisal versus Reading Time Relationship (Previous Word):** Non-linear GAMs are in green and linear control GAMs are in dotted blue. Results are for gaze duration. Shaded regions represent bootstrapped 95% confidence intervals. Gray subplots indicate the distribution of surprisal values.

We now briefly explain the components of these regressions. `s()` sets up a spline-based smooth term between a predictor and response variable that can take on a wide variety of non-linear functional relationships. Here, `k=6` indicates a maximum of 6 basis functions for the smooth. We choose `k=6` following the logic from Hoover et al. (2022), Appendix C. Having 6 basis functions allows for five degrees of freedom, which enables the regression to fit non-linear yet still relatively simple curves. The other term, `te()`, sets up a tensor product smooth term, which can effectively capture non-linear interactions between two variables.

B Surprisal versus RT for w_{t-1}

As mentioned in the main text, previous work has investigated the relationship between surprisal and reading times not just for the current word w_t, but also for the previous word, w_{t-1}. Looking at gaze duration in the Dundee corpus of English (Kennedy et al., 2003), Smith and Levy (2013) find an effect of w_{t-1}'s surprisal which is about as strong as the effect of w_t's surprisal on the reading time of w_t. In Figure 7 we show this relationship in our corpus for mGPT and monoT(all), using the same methods and presentational paradigm as in Section 5.1.

The results are consistent across models used, and suggest that the relationship between reading time and surprisal of the previous word is somewhat variable across languages. For English, Italian, Korean, Russian, and Spanish we find a relationship that is roughly linear and increasing, i.e., similar to the results for surprisal of the current word. For Dutch, Turkish, and Hebrew, we find a relationship that is roughly increasing, but visually non-linear. For Finnish, German, and Greek, we find either a flat or negative relationship. These results are in line with the effect terms plotted in Figure 3, where we find very weak and sometimes negative coefficients for the w_{t-1} surprisal term for these languages (i.e., the middle x-tick position in the top row). Overall, these results are consistent with the linear effect that has been previously observed in English. However, they suggest that the impact of the surprisal of the previous word varies between languages.