# The Linearity of the Effect of Surprisal on Reading Times across Languages

**Weijie Xu**
University of California, Irvine
weijie.xu@uci.edu

**Jason S. Chon**
Stanford University
jchon5@stanford.edu

**Tianran Liu**
University of Washington
tianranl@uw.edu

**Richard Futrell**
University of California, Irvine
rfutrell@uci.edu

## Abstract

In psycholinguistics, surprisal theory posits that the amount of online processing effort expended by a human comprehender per word positively correlates with the surprisal of that word given its preceding context. In addition to this overall correlation, more importantly, the specific quantitative form taken by the processing effort as a function of surprisal offers insights into the underlying cognitive mechanisms of language processing. Focusing on English, previous studies have looked into the linearity of surprisal on reading times. Here, we extend the investigation by examining eye-tracking corpora of seven languages: Danish, Dutch, English, German, Japanese, Mandarin, and Russian. We find evidence for superlinearity in some languages, but the results are highly sensitive to which language model is used to estimate surprisal.

## 1 Introduction

A great deal of insight into human language processing can be gleaned by studying the word-by-word processing difficulty experienced by a comprehender who is listening to or reading language. Consequently, the timecourse of incremental language processing difficulty has formed one of the major objects of study in psycholinguistics (Sedivy, 2014, Ch. 8). Here we take up the question of the functional form of the effect of **surprisal** (the negative log probability of each word given preceding context) on word-by-word processing difficulty, as indexed by reading time. Whereas previous empirical work has addressed this question in English, we study the question in the existing reading time databases of Danish, Dutch, English, German, Japanese, Mandarin Chinese, and Russian. To preview the results, we find consistent evidence for superlinearity for German and Russian; for all other languages except Mandarin, we find a linear effect of surprisal, but the analysis is inconclusive about the presence of additional superlinear effects

and the result is highly sensitive to which language model is used to estimate surprisal.

## 2 Background

In the field of psycholinguistics, **Surprisal Theory** (Hale, 2001; Levy, 2008) holds that processing difficulty is information-theoretic in nature. Letting $w_t$ be the word that a comprehender is perceiving at time $t$, and $w_{<t}$ be the sequence of all previous words, the surprisal $S_t$ measures the amount of information contained in $w_t$ in context:

$$S_t \equiv -\log p\left(w_t \mid w_{<t}\right). \qquad (1)$$

In general, Surprisal Theory holds that the amount of processing difficulty (for our purposes, the reading time, or RT) experienced by a comprehender for word $w_t$ in context $w_{<t}$ is given by $f(S_t)$, where $f$ is a monotonically increasing **linking function** converting bits of information into milliseconds of RT.

Surprisal Theory is well-supported empirically by findings of a correlation between surprisal and RT, and the ability of surprisal to explain RT effects previously attributed to other processing mechanisms (Levy, 2008), although it may not explain the full range of human processing difficulty (van Schijndel and Linzen, 2021; Arehalli et al., 2022). The correlation between surprisal and RT gets stronger when probabilities are calculated using lower-perplexity language models (Goodkind and Bicknell, 2018; Hao et al., 2020; Wilcox et al., 2020), although this relationship may not be universal across languages (Kuribayashi et al., 2021), and breaks down for recent large models (Oh et al., 2022; Oh and Schuler, 2023), whose ability to predict words is super-human (Buck et al., 2022).

The goal of this work is to empirically investigate the form of the linking function $f$ across languages. Although early work did not commit to a specific form for the linking function $f$ (Hale,

| Language | Corpus | Genre | Tokens | Subjects |
|---|---|---|---|---|
| Danish | Copenhagen Corpus (Hollenstein et al., 2022) | Speeches | 26,454 | 57 |
| Dutch | Ghent Eyetracking Corpus (GECO; Cop et al., 2017) | Novel | 58,302 | 18 |
| English | Dundee Corpus (Kennedy et al., 2003) | Editorials | 24,679 | 10 |
| German | Potsdam Sentence Corpus (Kliegl et al., 2006) | Artificial | 557 | 273 |
| Japanese | BCCWJ-EyeTrack (Asahara et al., 2016) | News | 970 | 24 |
| Mandarin | Beijing Sentence Corpus (Yan et al., 2010) | Artificial | 1,521 | 30 |
| Russian | Russian Sentence Corpus (Laurinavichyute et al., 2019) | Artificial | 892 | 96 |

Table 1: Eyetracking corpora used as datasets. 'Artificial' genre refers to sentences written by experimenters and presented in isolation. 'Tokens' indicates the number of text tokens (as defined using the tokenization provided by the eyetracking corpus) for which there is reading time data from at least one subject after all data exclusions.

2001; Levy, 2005), subsequent theoretical and empirical developments have suggested that the linking function $f$ might be a simple proportionality, that is $f(S_t) = k \cdot S_t$, where $k$ is a scalar conversion factor from bits of information to milliseconds of RT (Levy, 2008; Smith and Levy, 2013). Several studies have found evidence for linearity of the linking function $f$ using both $n$-gram models and modern language models (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Merkx and Frank, 2021; Aurnhammer and Frank, 2019; Wilcox et al., 2020; Shain et al., 2022).

However, the picture is complicated by theoretical considerations and recent empirical evidence. Theoretically, a superlinear function is favored by a model of incremental language processing in which an incremental parse tree is derived from a sentence prefix by sampling (Hoover et al., 2023), and the possibility of a superlinear linking function is related to the observation that speakers tend to communicate information at a relatively constant rate per unit time (Fenk and Fenk, 1980; Levy and Jaeger, 2007; Jaeger, 2010; Clark et al., 2023): such effects can be predicted from Surprisal Theory if processing cost is superlinear in surprisal (Smith and Levy, 2013; Levy, 2005, 2018). Empirically, Meister et al. (2021) find evidence for a small superlinear effect of surprisal on reading times at the level of sentences, and Hoover et al. (2023) find superlinear effects at the level of words.

Our work addresses the fact that broad-coverage psycholinguistic studies of surprisal have for the most part used reading time corpora of English only, and the functional form of the effect of surprisal on word RT in particular has only been investigated outside of English in one other study (Wilcox et al., 2023, to be discussed below in Section 5) to our knowledge.

## 3 Methods

### 3.1 Reading-time data

We use the reading-time datasets described in Table 1. There are two classes of reading-time corpus. The first is corpora of large amounts of naturalistic connected text, such as the English Dundee corpus, the Dutch GECO corpus, the Danish CopCo corpus, and the Japanese BCCWJ-EyeTrack; these corpora contain many different tokens in different contexts. The second class is the 'Sentence Corpora' of Mandarin, German, and Russian; these corpora contain only a few hundred artificially-constructed sentences which are read in isolation, but they have data from many participants.

Following Smith and Levy (2013), we examine first-pass gaze durations per word when this is possible based on the available eyetracking data; otherwise we examine first fixation duration. First-pass gaze duration, also called first-pass reading time and first-run dwell time, is defined as the sum of the duration of all fixations on a word $w$, starting with the first fixation on $w$ and ending with any fixation on any other word.[1]

**Data exclusions.** Following Smith and Levy (2013), we exclude the following datapoints from our analysis: words at the beginning or end of a line (when this could be determined in a corpus); words attached to punctuation[2]; and words that were skipped or have an RT of 0 ms. We also ex-

---

[1] For the English Dundee corpus, we used the sum of FDUR values in the first pass. For Dutch, we used the WORD_FIRST_FIXATION_TIME field. For German, we used the dur field. For Mandarin, we used the fixation_duration field. For Danish, we used the word_first_pass_dur field. For Japanese, we used the fpt field. For Russian, we used the IA_FIRST_DWELL_TIME field.

[2] This exclusion does not affect the Sentence Corpora, where words were presented without punctuation.

| Language | Train Size | Dev. PPL |
|----------|-----------:|---------:|
| Danish   | 42M        | 46.3     |
| Dutch    | 166M       | 32.9     |
| English  | 1,856M     | 32.6     |
| German   | 831M       | 31.8     |
| Japanese | 292M       | 52.2     |
| Mandarin | *804M*     | *3.1*    |
| Russian  | 451M       | 32.6     |

Table 2: Summary of the training set size and the dev set perplexity of the monolingual surprisal models trained in Wikipedia text. Training set sizes are per-token according to our trained tokenizers. The training set size and dev set perplexity for Mandarin are italicized to indicate that these are per-byte, rather than per-word, reflecting the byte-level tokenization of our Mandarin language model.

clude datapoints with a surprisal value greater than 20 nats or an RT greater than 2000 ms.

## 3.2 Language models

We use two sets of language models to derive surprisal estimates. First, we use the existing mGPT model (Shliazhko et al., 2022), a large multilingual model trained on Wikipedia text also used by Wilcox et al. (2023). Second, we train new autoregressive language models using the GPT-2 architecture (Radford et al., 2019) on text from Wiki40b dataset (Guo et al., 2020), removing title and section headers and splitting the remaining content by paragraph. We use the provided training/dev set splits. For all languages except Japanese and Mandarin, we train the default GPT-2 tokenizer on the training set. For Japanese, the GPT-2 tokenizer uses excessive memory, so we use the XLNet tokenizer instead (Yang et al., 2019). For Mandarin, we tokenize the text by byte (Xue et al., 2022), because the GPT-2 tokenizer splits up the sentences in the eyetracking corpus very differently from the provided tokenization.

We train language models to minimize cross-entropy loss on the training set using the `huggingface` library (Wolf et al., 2020). For languages other than Mandarin, we continue training until either dev set loss reaches 3.5 nats, or 7 epochs through the training data have been completed. For Mandarin, we continue training until the decrease in dev set loss is flat by visual inspection; the special treatment of Mandarin is because its byte-level loss is not comparable to the token-level loss for the other languages. Table 2 shows training set size and final language model perplexities.[3]

## 3.3 Statistical Analysis

We analyze the RT data by fitting linear mixed-effects models (Baayen et al., 2008) to predict RT as a function of surprisal, controlling for word length. In order to test for a linear effect of RT, we fit a sequence of models:

1. $M_0$: A regression including word length and log-transformed word frequency from Speer (2022) as a control predictor, with random intercepts by subject. This is the maximal consistently converging random effect structure (Barr et al., 2013).

2. $M_L$: $M_0$ plus a linear effect of surprisal.

3. $M_{LQ}$: $M_L$ plus a quadratic effect of surprisal, testing for a superlinear effect beyond the linear effect.

The existence of a linear effect was determined by a likelihood ratio test comparing $M_0$ to $M_L$. The existence of an additional superlinear effect was tested by comparing $M_L$ to $M_{LQ}$.

**Spillover.** In reading time data, processing slowdown due to factors such as surprisal is often delayed, appearing on following words (Erlich and Rayner, 1983). In order to control for this effect, we adopt the standard practice in psycholinguistics: for all the variables in our models, we use not only the value of the current word, but also the value of the $K$ previous words, where the value of the spillover window size $K$ is determined by the following procedure: first a control model is fit to the reading time data using only word length as a predictor, and then a second model is fit using the same variables, incrementing the spillover window size by one. If the incremented model is a significantly better fit to the data than the control model by a likelihood ratio test, then the size of the spillover window is expanded. The resulting spillover window sizes for the four corpora are shown in Table 3. If a token does not have valid surprisal values for all previous tokens within its spillover window, it is excluded from statistical analysis. Results following an alternative procedure where the spillover window is fixed at 2 for all corpora are shown in Appendix A.

---

[3] The trained LMs are available at https://huggingface.co/rfutrell.

| | | mGPT | | | Monolingual | | |
|---|---|---|---|---|---|---|---|
| Language | Spillover | ET PPL | Linear | Quadratic | ET PPL | Linear | Quadratic |
| Danish | 3 | 41.1 | + | $p = 0.427$ | 76.7 | + | + |
| Dutch | 3 | 124.5 | + | − | 295.7 | + | + |
| English | 2 | 52.0 | + | $p = 0.935$ | 74.9 | + | + |
| German | 3 | 37.1 | + | + | 186.9 | + | + |
| Japanese | 0 | 16.1 | + | $p = 0.354$ | 215.9 | + | $p = 0.326$ |
| Mandarin | 0 | 21.4 | $p = 0.275$ | $p = 0.487$ | *30.7* | $p = 0.284$ | $p = 0.081$ |
| Russian | 3 | 27.1 | + | + | 220.6 | + | + |

Table 3: Summary of results using the mGPT language model and the monolingual models trained in Wikipedia text to estimate surprisal. 'Spillover' is the spillover window size determined by the procedure in Section 3.3. 'ET PPL' is the perplexity of each language model evaluated on the eyetracking corpus text. Note that the perplexities across models are not directly comparable because of different tokenizations. The monolingual perplexity for Mandarin is italicized to indicate that these are per-byte, rather than per-word, reflecting the byte-level tokenization of our Mandarin language model. 'Linear' is the sign of any significant (at $p < 0.05$) linear effect of surprisal on RT for the current token, and 'Quadratic' is the sign of any significant quadratic effect for the current token; $p$ values are presented if the effect is not significant.

## 4 Results

Figure 1 shows average reading time as a function of surprisal, along with linear fits, using the mGPT surprisal estimates. A strong positive relationship always obtains. Visually, the effect appears to be largely linear in the larger corpora of Danish, Dutch, and English, but the shape of the effect is harder to discern for the the smaller corpora (German, Japanese, Mandarin, and Russian).

Regression results for both surprisal models are schematized in Table 3, which gives the sign of the coefficient for (linear or quadratic) surprisal in the fitted regression predicting reading time. As expected, the sign of the linear effect is always positive, meaning that increasing surprisal corresponds to increasing reading time; the effect is significant in all languages except Mandarin. For the mGPT model, the quadratic effect is significantly negative in Dutch, indicating sub-linearity, and significantly positive in German and Russian, indicating superlinearity. The quadratic effect is not significant elsewhere. For the monolingual models, we find positive quadratic effects in all languages except Japanese and Mandarin.

The full regression code and results are available at `https://github.com/weijiexu-charlie/Linearity-of-surprisal-on-RT`.

### 4.1 Discussion

Overall, the results provide evidence for a positive linear effect of surprisal on reading time across languages, and inconsistent evidence for a superlinear effect: the quadratic effect of surprisal varies in significance and sign depending on the language model used.

It is hard to discern a systematic difference between languages showing evidence for only a linear effect and those showing possible deviations from linearity—a consistent superlinear effect is found only in German and Russian, the two smallest datasets. The evidence for superlinearity does not seem to follow a typological pattern.

## 5 Concurrent work

Concurrent to our work, Wilcox et al. (2023) have examined the linearity of the effect of surprisal on RT in 11 languages of the MECO eyetracking corpus (Siegelman et al., 2022), finding consistent evidence for linearity and a lack of evidence for superlinearity across all languages studied.[4] Our work is complementary with this other work, differing in datasets, languages considered, and statistical methodology. Regarding data, Wilcox et al. (2023) use the MECO corpus, consisting of parallel translated text, which is more similar in size to the 'Sentence Corpora' that we use for German, Mandarin, and Russian. In contrast, we use larger corpora of naturalistic text for Danish, Dutch, English, and Japanese. We also investigate reading times in Japanese and Mandarin, which

---

[4]Another related work is de Varda and Marelli (2022, 2023), who study the existence of a surprisal effect across languages; however, they do not test the functional form of surprisal.
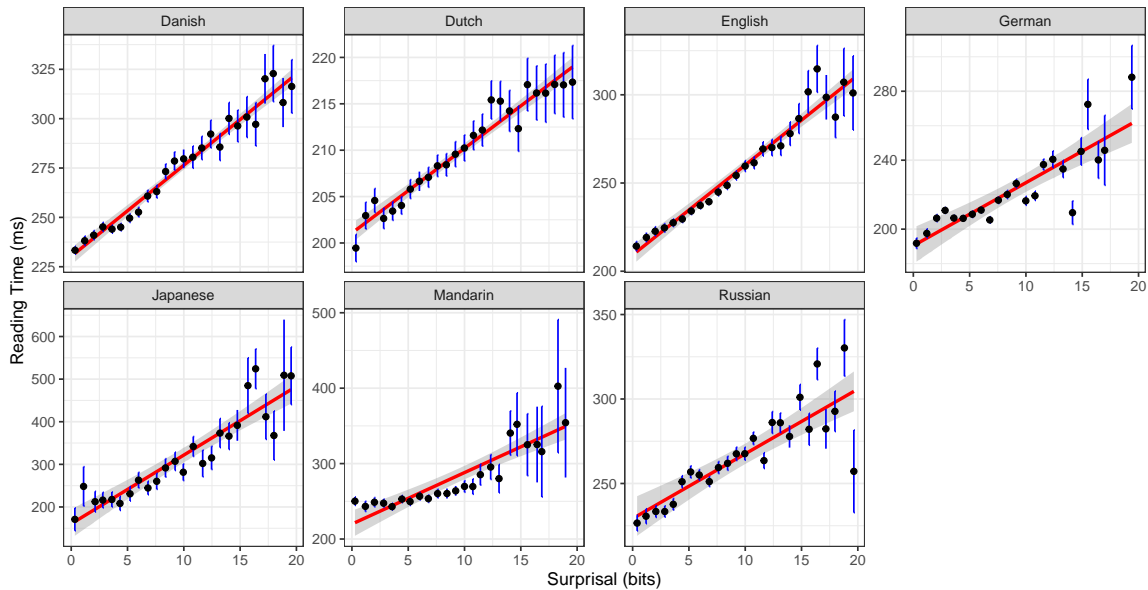
Figure 1: Average RT as a function of the surprisal from mGPT. Surprisal is binned into 25 categories, and the mean RT within each category is shown in black with a 95% confidence interval in blue. A linear fit to these points is shown in red.

are not present in Wilcox et al. (2023). Reading times in these languages are especially interesting because they are written using orthographies that do not mark word boundaries; thus reading in these languages involves some arguably different cognitive processes than the other languages, as word/morpheme segmentation must be inferred from the text.

Our work also differs from Wilcox et al. (2023) in the statistical methodology. Following Smith and Levy (2013) and Wilcox et al. (2020), Wilcox et al. (2023) compare fits of linear and non-linear Generalized Additive Models when predicting reading times from surprisal after controlling for word length and frequency, with a fixed spillover window across languages. In contrast, we adopt a simpler approach: we compare a mixed-effects linear model against a quadratic model, with a variable spillover window across languages.

Resolving the apparent tension between our results, which join Hoover et al. (2023) in finding possible evidence for superlinearity, against the results suggesting pure linearity (Shain et al., 2022; Wilcox et al., 2023, for example) should be a focus in future research.

## 6 Conclusion

We conclude that the effect of surprisal on reading times across languages may be slightly superlinear, but the evidence for this superlinearity is unstable

and depends on the language model used. Determining the functional form of the effect in further languages will require reading time corpora of large amounts of naturalistic text and strong language models.

## Limitations

The generality of our conclusions is limited by factors related to statistical methodology, the nature of the corpora used, and the nature of the language models trained.

**Statistical Methodology**   Our analysis is limited because we adopt a relatively simple statistical methodology, in three ways.

First, our handling of spillover effects, while standard in the psycholinguistic literature, does not account for the full shape of the timecourse of the effect of surprisal, which may lead to a lack of power to detect effects (Shain and Schuler, 2018). More sophisticated methods for handling spillover effects, such as deconvolutional time series methods (Shain, 2019; Shain et al., 2022), may yield different results. The procedure of setting a fixed spillover window also limits statistical power because a token can only be kept in the regression if it has valid surprisal values for all previous tokens within the spillover window; the result is that many tokens must be thrown out.

Second, our test for superlinear effects of sur-

prisal on RT is relatively simple: we compare a model with only a linear effect against a model with a linear+quadratic effect. In contrast, Smith and Levy (2013) search for linearity using a method involving cubic splines, and Meister et al. (2021) fit models predicting RT as

$$\text{RT} = k \cdot S_t^{\alpha} \qquad (2)$$

for varying $\alpha$, allowing them to quantify the shape of the superlinearity (they find best overall fits with values between $\alpha = 1$ and $\alpha = 1.5$). Our method can detect the existence of a superlinear effect but cannot find a specific exponent $\alpha$ to characterize it.

Third, we use linear regression to predict reading times, because we want to study a linear effect of surprisal on RT in particular. However, linear regression assumes a Gaussian error distribution, which does not generally hold for human reaction times: human reaction times cannot be negative and tend to be skewed to the right, and have been characterized using distributions such as Log-Normal (Baayen and Milin, 2010) and Ex-Gaussian (Luce, 1986). The use of a more realistic error distribution may lead to different results, especially in our critical datapoints where RT values are higher than expected from linear Surprisal Theory.

**Eyetracking Corpora** We use publically-available eyetracking datasets which are not matched across languages, because the datasets were developed by different researchers for different purposes at different times. As a result, the corpora differ not only in language but also genre.

**Language Models** We only use one language model architecture (GPT), which is a potential source of inaccuracy in our surprisal values: other language model architectures may deliver surprisal values that better reflect human expectations. Furthermore, the ET corpus texts differ in genre from the Wikipedia text corpora used to train the language models. This discrepancy may result in surprisal values for the ET corpus texts that do not reflect human expectations. This factor may account for the discrepancy where we find a superlinear effect in English whereas some previous work has not (for example, Wilcox et al., 2020).

Another potential limitation is that we train all the monolingual language models to the same target perplexity, which may not reflect the same level of language model quality per language because different languages may have different inherent entropy rates (Bentz et al., 2017).

## Ethics Statement

We do not foresee ethical issues with this work.

## Acknowledgements

## 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.

Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. 2016. Reading-time annotations for "Balanced Corpus of Contemporary Written Japanese". In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 684–694, Osaka, Japan. The COLING 2016 Organizing Committee.

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

R. Harald Baayen, D.J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

R. Harald Baayen and Petar Milin. 2010. Analyzing reaction times. *International Journal of Psychological Research*, 3(2):12–28.

Dale J. Barr, Roger P. 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.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i-Cancho. 2017. The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, 19:275–307.

Buck, Fabien Roger, and LawrenceC. 2022. Language models seem to be much better than humans at next-token prediction. *LessWrong*.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger P. Levy. 2023. A Cross-Linguistic Pressure for Uniform Information Density in Word Order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

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.

Andrea de Varda and Marco Marelli. 2023. Scaling in cognitive modelling: a multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149, Toronto, Canada. Association for Computational Linguistics.

Kate Erlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning & Verbal Behavior*, 22:75–87.

August Fenk and Gertrud Fenk. 1980. Konstanz im Kurzzeitgedächtnis—Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie*, 27(3):400–414.

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.

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*.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86, Online. Association for Computational Linguistics.

Nora Hollenstein, Maria Barrett, and Marina Björnsdóttir. 2022. The Copenhagen Corpus of eye tracking recordings from natural reading of Danish texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France. European Language Resources Association.

Jacob Louis Hoover, Morgan Sonderegger, Steven T. Piantadosi, and Timothy J. O'Donnell. 2023. The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing. *Open Mind*, 7:350–391.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee corpus. Poster presented at the 12th European Conference on Eye Movement.

Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, 135(1):12.

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.

Anna K. Laurinavichyute, Irina A. Sekerina, Svetlana Alexeeva, Kristine Bagdasaryan, and Reinhold Kliegl. 2019. Russian sentence corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods*, 51(3):1161–1178.

Roger P. Levy. 2005. *Probabilistic Models of Word Order and Syntactic Discontinuity*. Ph.D. thesis, Stanford University, Stanford, CA.

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

Roger P. Levy. 2018. Communicative efficiency, uniform information density, and the Rational Speech Act theory. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, page 684–689.

Roger P. Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19, pages 849–856.

R. Duncan Luce. 1986. *Response Times*. Oxford University Press, New York.

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.

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

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5.

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.

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

Julie Sedivy. 2014. *Language in Mind: An Introduction to Psycholinguistics*. Oxford University Press.

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, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time. *PsyArXiv*.

Cory Shain and William Schuler. 2018. Deconvolutional Time Series Regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2679–2689, Brussels, Belgium. Association for Computational Linguistics.

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, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior Research Methods*, 54(6):2843–2863.

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

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

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

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 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *arXiv preprint arXiv:2307.03667*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Ming Yan, Reinhold Kliegl, Eike M. Richter, Antje Nuthmann, and Hua Shu. 2010. Flexible saccade-target selection in Chinese reading. *The Quarterly Journal of Experimental Psychology*, 63(4):705–725.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*.

# A  Results with a Fixed Spillover Window

Table 4 shows our main results but using a fixed spillover window of size 2 for all languages, for the mGPT and monolingual surprisal models. When using mGPT, fixing the spillover window yields a significant positive linear effect of surprisal in Mandarin, eliminates the evidence for the apparent sublinear effect in Dutch, and gives evidence for a superlinear effect in Japanese. When using the monolingual models, the results are the same as in the main text, but now with an apparent positive quadratic effect in Japanese.

| Language | Linear | | Quadratic | |
| --- | --- | --- | --- | --- |
| | mGPT | Monoling | mGPT | Monoling |
| Danish | + | + | ns, $p = 0.289$ | + |
| Dutch | + | + | ns, $p = 0.101$ | + |
| English | + | + | ns, $p = 0.935$ | + |
| German | + | + | + | + |
| Japanese | + | + | + | + |
| Mandarin | + | ns, $p = 0.058$ | ns, $p = 0.855$ | ns, $p = 0.424$ |
| Russian | + | + | + | + |

Table 4: Main results as in Table 3 using the mGPT surprisal model with afixed spillover window of size 2.

## B   Additional Figures

Figure 2 shows reading times as a function of surprisal based on the monolingual surprisal models. Figures 3 and 4 show residual reading times after controlling for spilled-over length and frequency, as a function of surprisal, for the mGPT and monolingual language models respectively.
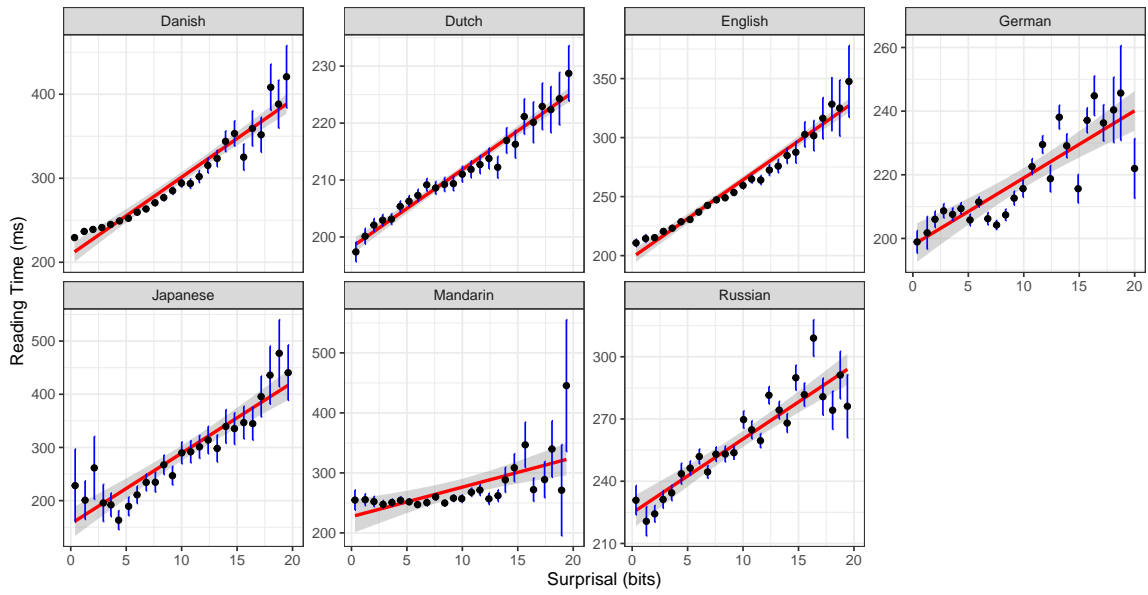
Figure 2: As in Figure 1, average RT as a function of the surprisal from monolingual LM.
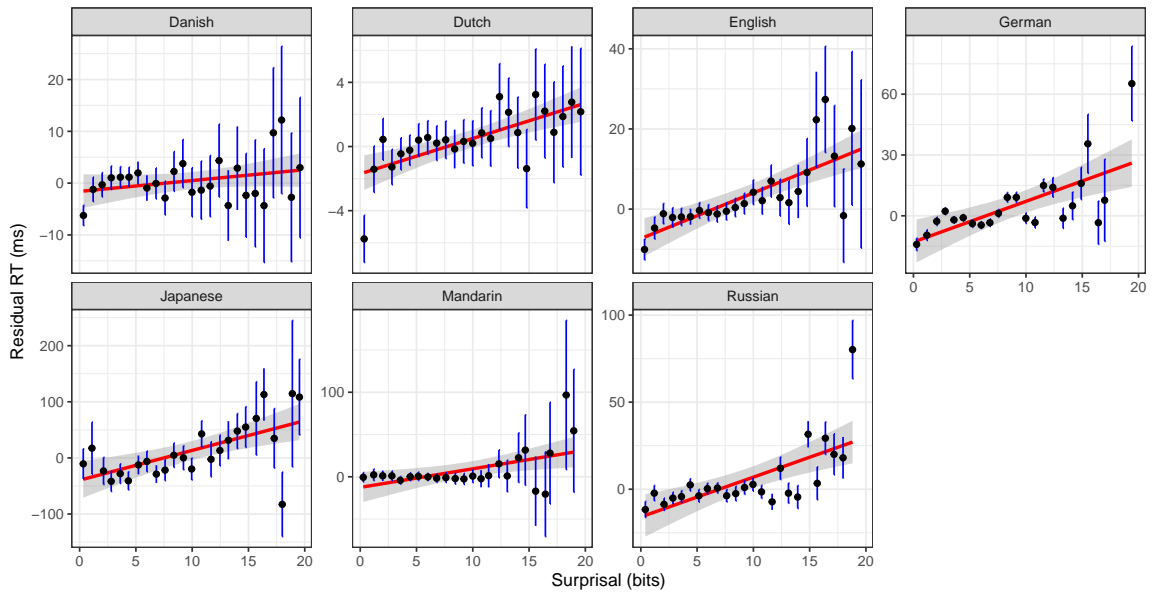


Figure 3: Average residual RT controlled by word length and word frequency as a function of the surprisal from mGPT.
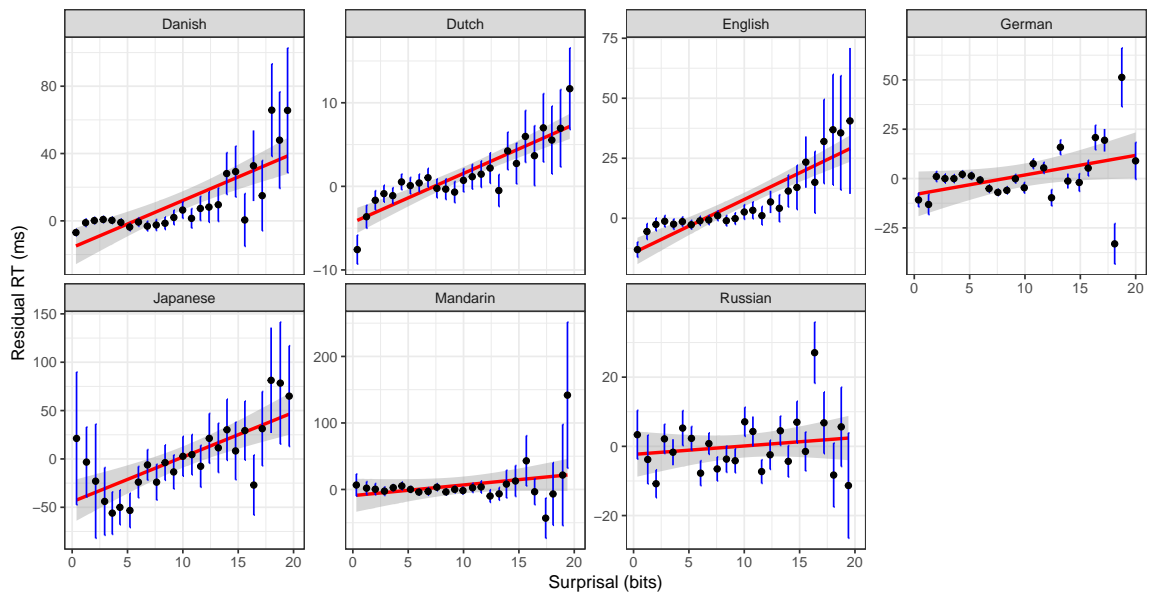
Figure 4: Average residual RT controlled by word length and word frequency as a function of the surprisal from monolingual LM.