The Impact of Token Granularity on the Predictive Power of Language Model Surprisal

Byung-Doh Oh
Center for Data Science
New York University
oh.b@nyu.edu

William Schuler
Department of Linguistics
The Ohio State University
schuler.77@osu.edu

Abstract

Word-by-word language model surprisal is often used to model the incremental processing of human readers, which raises questions about how various choices in language modeling influence its predictive power. One factor that has been overlooked in cognitive modeling is the granularity of subword tokens, which explicitly encodes information about word length and frequency, and ultimately influences the quality of vector representations that are learned. This paper presents experiments that manipulate the token granularity and evaluate its impact on the ability of surprisal to account for processing difficulty of naturalistic text and garden-path constructions. Experiments with naturalistic reading times reveal a substantial influence of token granularity on surprisal, with tokens defined by a vocabulary size of 8,000 resulting in surprisal that is most predictive. In contrast, on garden-path constructions, language models trained on coarser-grained tokens generally assigned higher surprisal to critical regions, suggesting a greater sensitivity to garden-path effects than previously reported. Taken together, these results suggest a large role of token granularity on the quality of language model surprisal for cognitive modeling.

1 Introduction

In cognitive modeling, word-by-word surprisal is often used as a predictor of processing difficulty, under a theoretical framework that emphasizes the predictive aspect of real-time language processing (Hale, 2001; Levy, 2008). In recent years, neural network-based language models (LMs) have been used to calculate and evaluate surprisal against human reading times (Wilcox et al., 2020; Merkx and Frank, 2021), which has opened possibilities for refining them as computational models of language processing and using them to study how predictive processing interacts with other cognitive processes. Therefore, a core question in this area is how various aspects of language modeling such as the LMs' architecture or training data influence the learned probabilities and their alignment to human-like processing difficulty.

'If you were to journey'

Finer granularity, more character-like ($|V| = 256$)
`␣` `I` `f` `␣` `y` `o` `u` `␣w` `er` `e` `␣to` `␣` `j` `o` `ur` `n` `e` `y`

Coarser granularity, more word-like ($|V| = 128000$)
`␣If` `␣you` `␣were` `␣to` `␣journey`

Figure 1: Smaller subword vocabulary sizes result in longer sequences of finer-granularity `tokens` that are more character-like (top), and larger vocabulary sizes result in shorter sequences of coarser-granularity `tokens` that are more word-like (bottom).

One such variable that has been overlooked in cognitive modeling is the granularity of the tokens over which LMs are trained to define probability distributions. To allow LMs to flexibly handle unseen word forms and keep the vocabulary size manageable, it has become a standard practice in language modeling to use 'subword' tokenizers (e.g. Sennrich et al., 2016; Kudo, 2018). Such tokenizers are often trained on corpus statistics such that their vocabularies contain a fixed number of frequent sequences (which may or may not correspond to words) as independent tokens. As a consequence, less frequent word forms are split into multiple subword tokens during both training and inference.

This suggests that there are at least two ways in which token granularity will influence the quality of LM surprisal in the context of cognitive modeling. The first is through the different initial biases about word probabilities that the different levels of token granularity embody. For example, a tokenizer with a very fine token granularity (Figure 1, top) tokenizes the word *journey* into seven tokens, and therefore a uniform initial distribution over tokens predicts its probability to be six magnitudes lower than that of *to*. In contrast, a tokenizer with coarser granularity (Figure 1, bottom) keeps both

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4150–4162
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

journey and *to* intact, and a uniform initial distribution over these tokens predicts their probabilities to be equal. As such subword tokenization is informed by word length and frequency, which are variables that are well known to influence real-time processing (Barton et al., 2014; Just and Carpenter, 1980), some tokenization schemes are more likely to lead to word-level surprisal that is more predictive of processing difficulty.

The second way in which token granularity will impact LM surprisal is through the quality of the token representations that are learned. That is, LMs trained on coarse-grained tokens that are more word-like will learn token representations that align closer to lexical co-occurrence statistics like those from earlier neural network LMs (Mikolov et al., 2013). In contrast, fine-grained tokens will cause words to be split across multiple vector representations, which may make learning word-to-word associations more challenging. For instance, the LM would need to attend to seven separate vector representations for *journey* to predict *travel* later in the sequence.

Motivated by these observations, this work presents experiments that manipulate the token granularity of LMs and evaluate its influence on the ability of surprisal to account for processing difficulty of both naturalistic text and garden-path constructions. First, regression experiments across multiple reading time corpora reveal a strong influence of token granularity on the predictive power of surprisal prior to any LM training, with tokens defined by a vocabulary size of 4,000 resulting in surprisal that is most predictive. This initial bias appears to persist throughout training in smaller LMs, and eventually yield surprisal that is more predictive than that of larger LMs. In contrast, LMs trained with tokens of coarser granularity generally assigned higher surprisal to critical regions of garden-path constructions, suggesting a greater sensitivity to garden-path effects. These results suggest a large role of token granularity on the quality of LM surprisal for cognitive modeling, with different levels of granularity being more appropriate for modeling broad-coverage comprehension and garden-path effects.

2 Related Work

The choice of subword tokenizers has mostly been studied in terms of performance on natural language processing tasks (e.g. Bostrom and Durrett, 2020), with more recent work aiming to provide explanations for their downstream impact (Zouhar et al., 2023; Schmidt et al., 2024) or the discrepancy in performance across languages (Arnett and Bergen, 2025). Other work proposes methods for mitigating 'quirks' introduced by subword tokenization, such as LMs' sensitivity to misspelling (Vieira et al., 2024) or 'under-trained' tokens with poor quality representations (Land and Bartolo, 2024).

In psycholinguistic modeling, Oh et al. (2021) incorporate a character model to an incremental left-corner parser and show improvements in surprisal's fit to naturalistic reading times. Nair and Resnik (2023) compare a subword tokenization scheme informed by concatenative morphology against a word-level and a byte-pair encoding scheme and report its impact on LM surprisal's predictions of reading times. Giulianelli et al. (2024) apply Vieira et al.'s (2024) algorithm to derive character-level probabilities from GPT-2's (Radford et al., 2019) token-level probabilities and demonstrate the potential of using them to model the processing difficulty of specific focal areas (e.g. the first three characters of a word) more flexibly.

3 Experiment 1: Impact on Fit to Naturalistic Reading Times

The first experiment evaluates the effect of token granularity on the fit of LM surprisal to naturalistic reading times. LM surprisal is first evaluated prior to any training to examine the extent to which word-level surprisal determined purely by token granularity is predictive of reading times. Subsequently, surprisal from LMs of different sizes is evaluated after training to examine how model size and training data interact with these initial biases.

3.1 Vocabulary Construction

We manipulate the granularity of the LM tokens by training subword tokenizers with vocabularies of different sizes, upon which the LM training and surprisal calculation is based.

Subword Tokenizer. We used the unigram language model tokenizer (ULM; Kudo, 2018) to construct subword vocabularies of different sizes. The ULM tokenizer aims to find a vocabulary V of a desired size that maximizes the joint probability of the subword sequence. This is achieved by starting with a large vocabulary and then iteratively pruning entries that result in the smallest drop in

the marginal likelihood over possible tokenizations. The ULM tokenizer is similar to the byte-pair encoding tokenizer (BPE; Sennrich et al., 2016) in that it 'compresses' frequent subword sequences into tokens, but different in that it treats the character as the basic subword unit (cf. bytes as basic units in BPE[1]) and defines a probability distribution over possible tokenizations given a string (cf. BPE is deterministic).

Training Data. The ULM tokenizer requires a training corpus for estimating the probability of each subword token and ultimately constructing the vocabulary. To this end, we used a subset of the English training section of the Wiki-40B dataset (Guo et al., 2020), which contains Wikipedia articles. The Wiki-40B dataset was chosen as it closely approximates the domain represented by the reading time corpora that are of interest, in contrast to other pre-training corpora that also include e.g. programming code. After removing the metadata tags and filtering articles that are not in English, 1,000,000 articles were sampled to train the ULM tokenizer.

Training Setup. Subsequently, ULM tokenizers were trained to vocabulary sizes of {256, 512, 1000, 2000, 4000, 8000, 16000, 32000, 48000, 64000, 128000} to cover a wide range of token granularity.[2] The character-level coverage of the training corpus was set to 99.95%, which resulted in a set of 158 characters as the basic subword units. The final vocabulary of each condition consists of these basic subword units, frequent subword sequences determined by the training process, an <unk> token for characters that are not in the set of basic subword units, and an <s> token for denoting the start of sequence.

3.2 Language Modeling

Based on the tokenizers trained following the procedures in Section 3.1, autoregressive LMs with different vocabulary sizes were subsequently trained.

Model Architecture. As can be seen in Figure 1, different token granularities result in token sequences of varying lengths over the same text string. This poses a stark challenge for training Transformers (Vaswani et al., 2017) with different vocabulary sizes, whose self-attention mechanism has unfavorable space complexity for long input sequences.[3] To overcome this issue, we train LMs based on the Mamba-2 architecture (Dao and Gu, 2024), which belongs to the class of state space models (SSMs). Each Mamba-2 block aggregates representations from previous timesteps through recurrence-like operations defined by timestep-specific parameters that are determined by the input. Similarly to Transformer's attention, Mamba-2's operations can be distributed across multiple 'heads' that have different parameters. For the experiments described in this paper, we use the multi-input setup of Mamba-2, which was shown to be the most effective for language modeling (Dao and Gu, 2024).[4]

Model	#L	#H	d_{model}	#Parameters
Small	6	8	256	2,592,400
Medium	12	16	512	19,847,744
Large	24	24	768	87,993,792

Table 1: Hyperparameters of LMs that were trained in this work. #L, #H, and d_{model} refer to number of Mamba-2 layers, number of SSM heads per layer, and model embedding size, respectively. #Parameters exclude parameters from the token embeddings, the number of which are different across each tokenizer.

To examine the potential impact of the LM's size on the fit of surprisal to reading times (Oh and Schuler, 2023b; Shain et al., 2024), we trained models of three different sizes for each tokenizer condition. Following conventional practice, we varied the number of layers,[5] the number of 'SSM heads' on each layer, and the model embedding size. Additionally, following Dao and Gu (2024), the state size was set to twice the model embedding size, and the size of each SSM head was fixed to 64. Finally, the token embeddings were shared across the initial input and the final projection layers. The resulting hyperparameters and the total number of non-embedding parameters are summarized in Table 1.

[1] As bytes are less interpretable than characters, the ULM tokenizer was opted for in this work.

[2] LMs widely used in cognitive modeling like GPT-2 (Radford et al., 2019) or Pythia (Biderman et al., 2023) have a vocabulary size of around 50,000.

[3] The conventional solution to this issue is to set a maximum input sequence length, which would nonetheless result in LMs that condition on different amounts of text depending on token granularity.

[4] The multi-input setup is analogous to multi-value attention, where different attention heads share the query and key matrices but not the value matrices.

[5] Because a Mamba-2 layer does not contain a multi-layer perceptron like a Transformer layer, Dao and Gu (2024) argue that a Mamba-2 model needs twice as many layers in order to be comparable with a Transformer model.

Training Data. We used the entire English training section of the Wiki-40B dataset (Guo et al., 2020) to train the LMs. Following similar procedures as the tokenizer training, metadata tags and articles that are not in English were removed. Moreover, one article that overlapped substantially in content with the reading time corpora was also filtered out. Each remaining article was then treated as a single training example for the LMs. The ULM tokenizers trained in Section 3.1 were used to tokenize the articles by returning the most likely tokenization of the string.

While the Mamba-2 architecture is more favorable toward longer sequences, some articles were excessively long when tokenized into finer-grained tokens. Therefore, a small subset of long articles was further split into 'Wikipedia sections' to alleviate this issue. This resulted in a total of 5,152,219 training examples.

Training Setup. One iteration of the training data was provided in the same order to each LM in 10,063 training batches of 512 examples. The AdamW optimizer (Loshchilov and Hutter, 2019) with a maximum learning rate of 1×10^{-3} was used to train the model parameters. This maximum learning rate was linearly warmed up over the first ~5% of training steps (i.e. 503 steps) and was subsequently annealed to a minimum of 1×10^{-5} following a cosine schedule over the remaining training steps. Gradients were clipped to a maximum norm of 1 to further stabilize training. All LM training took place in half-precision on a 48GB Nvidia RTX 8000 GPU.[6]

3.3 Reading Time Modeling

Surprisal from LMs trained following the procedures in Section 3.2 was subsequently evaluated on its ability to predict naturalistic reading times. Following recent work (Shain, 2024; Shain et al., 2024), this experiment aimed to identify trends using data across multiple reading time corpora.

Reading Time Corpora. The reading time data analyzed in this experiment consist of 10 measures from five self-paced reading (SPR) and eye-tracking (ET) corpora:

1. Natural Stories (Futrell et al., 2021): SPR times from 181 subjects that read 10 naturalistic English passages of narrative and expository text (10,256 words).
2. Brown (Smith and Levy, 2013): SPR times from 35 subjects that read 13 English passages (7,180 words) from the Brown Corpus (Kučera and Francis, 1967).
3. GECO (Cop et al., 2017): Fixation durations from 14 monolingual subjects that read the English version of novel *The Mysterious Affair at Styles* (Christie, 1920; 13 chapters, 56,411 words).
4. Dundee (Kennedy et al., 2003): Fixation durations from 10 subjects that read 67 English newspaper editorials (51,501 words).
5. Provo (Luke and Christianson, 2018): Fixation durations from 84 subjects that read 55 short English passages (2,746 words) ranging between news articles, science magazines, and fictional work.

Corpus/Measure	Fit	Exploratory	Held-out
Natural Stories SPR	384,984	192,826	192,449
Brown SPR	59,292	29,671	30,157
GECO FP	144,850	72,468	72,574
GECO GP	144,850	72,468	72,574
Dundee SP	155,483	77,809	77,101
Dundee FP	98,115	48,598	48,794
Dundee GP	98,115	48,598	48,794
Provo SP	91,032	45,654	45,404
Provo FP	52,959	26,539	26,640
Provo GP	52,960	26,539	26,640

Table 2: Number of data points in each partition of each reading time dataset.

Data Preprocessing and Partitioning. For the SPR datasets, by-word reading times were filtered to exclude those of words at sentence boundaries (i.e. sentence-initial and -final) and those shorter than 100 ms or longer than 3,000 ms. For the Natural Stories Corpus, data from subjects who answered fewer than five comprehension questions correctly were also removed.

For the ET data that contains non-linear eye movements, the scan path (SP), first-pass (FP), and go-past (GP) durations were calculated and analyzed for each word region.[7] These datasets were filtered to remove data for unfixated words, words following saccades longer than four words, and words at sentence and document boundaries. Data

[6]The trained ULM tokenizers and Mamba-2 models are publicly available at https://github.com/byungdoh/ssm-surprisal.

[7]The SP duration could not be calculated for the GECO dataset that does not provide raw fixation data.

points corresponding to words at line and screen boundaries were also excluded for the Dundee Corpus that provides relevant annotations.

After data preprocessing, each dataset was partitioned into fit, exploratory, and held-out partitions that comprise roughly 50%, 25%, and 25% of data points respectively. This partitioning was conducted based on the sum of the subject ID and the sentence ID,[8] which keeps all data from a particular subject reading a particular sentence in one partition. The fit partition was used to estimate the regression parameters, and all results are reported on the exploratory partition. The held-out partition is reserved for any statistical significance testing, and its use is kept minimal. The number of data points after preprocessing and partitioning is outlined in Table 2.

Surprisal Calculation. Each passage of the five reading time corpora was tokenized using each LM's respective ULM tokenizer and provided as input to calculate token probabilities that were converted to word probabilities. Preliminary analyses showed that the trained ULM tokenizers prepend the whitespace character to tokens, such that they have leading whitespaces. However, if word probabilities are naively calculated with leading whitespaces (e.g. P(*were* | *If you*) calculated as P(␣were | ␣If ␣you) in Figure 1), the sum over all word probabilities can exceed one, as the tokens do not explicitly mark the end of the word. Recent work (Oh and Schuler, 2024; Pimentel and Meister, 2024) provides a correction for this issue that factors the probability of each whitespace and re-allocates it to its preceding token (i.e. P(*were* | *If you*) calculated as P(were ␣ | ␣If ␣you ␣)), which was applied in this work to calculate word probabilities.

Regression Modeling. We fit linear mixed-effects (LME; Bates et al., 2015) regression models using each fit partition of the reading time datasets to evaluate the impact of token granularity on LM surprisal's fit to reading times. The goodness-of-fit was evaluated by calculating the increase in regression model log-likelihood (ΔLogLik) on each exploratory partition due to including each LM surprisal predictor on top of the baseline regression model.

The baseline regression models contain as predictors word length in characters, index of word position within the sentence, unigram surprisal (all datasets), and whether the previous word was fixated (ET datasets only). Unigram surprisal was calculated using the KenLM toolkit (Heafield et al., 2013) with parameters estimated on the OpenWebText Corpus (Gokaslan and Cohen, 2019). On top of these baseline regression models, surprisal of the current word and the preceding word was included to capture spillover effects (Rayner et al., 1983).

The raw reading times were not transformed prior to regression modeling, assuming a linear relationship between surprisal and reading times (Wilcox et al., 2023b; Xu et al., 2023; Shain et al., 2024). The LME models were fit with maximal random effects that were supported by the data (Barr et al., 2013) by removing the least predictive random effect until all LME models converged. The models fit to SPR data included by-subject random slopes for word position, word length, and surprisal of current and previous word. The models fit to ET data included by-subject random slopes for word position and surprisal of current word. All LME models also include a by-subject random intercept. These regression modeling procedures were repeated for all LMs prior to any LM training (i.e. at initialization) and at the end of LM training. The corpus-level perplexity is also reported at the end of LM training, based on prior results showing a systematic relationship between perplexity and surprisal's fit to reading times (Wilcox et al., 2020, 2023a; Oh and Schuler, 2023a).

3.4 Results

The results from LMs prior to training in Table 3 reveal a strong influence of token granularity on the predictive power of word-level surprisal, with surprisal from tokenizers with vocabulary sizes of 4,000, 8,000, and 16,000 showing the strongest fits. As the vocabulary size increases and subword tokens become more word-like, the fit of surprisal to reading times declines as more and more words are similarly assigned 'uniform surprisal.' On the other hand, as the vocabulary size decreases and subword tokens become more character-like, the fit to reading times also declines because surprisal becomes more strongly correlated with word length that is included as a baseline predictor. The intermediate vocabulary sizes appear to represent an optimum along this continuum.

At the end of LM training, a strong interaction is observed between LM size and token granularity (Figure 2). While surprisal from the *Small* LMs

[8] To this end, the sum of the subject ID and sentence ID modulo four was calculated (zero or one: fit partition, two: exploratory partition, three: held-out partition).

Model \ $\lvert V \rvert$	256	512	1000	2000	4000	8000	16000	32000	48000	64000	128000
Small	2205.8	2292.7	2338.2	2386.2	**2560.2**	2531.5	2552.0	2360.3	2220.8	2088.6	1901.9
Medium	2237.5	2312.7	2344.8	2393.5	**2555.8**	2541.0	2533.7	2360.5	2210.9	2101.5	1902.9
Large	2212.1	2274.4	2321.6	2395.2	**2542.4**	2519.0	2536.2	2362.8	2204.1	2049.3	1890.9
Average	2218.5	2293.3	2334.9	2391.6	**2552.8**	2530.5	2540.6	2361.2	2211.9	2079.8	1898.6

Table 3: ΔLogLik from LM surprisal prior to any LM training, aggregated over the exploratory partitions of all reading time datasets.

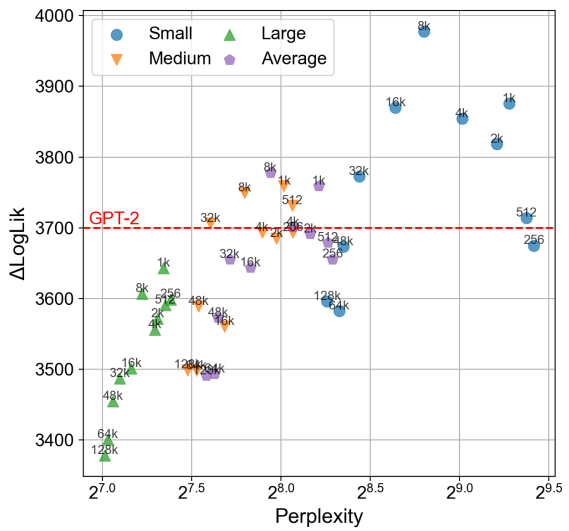

Figure 2: ΔLogLik from LM surprisal on the exploratory partitions and corpus-level perplexity after LM training, both aggregated over all reading time corpora. The 'Average' represents the arithmetic mean of ΔLogLik and the geometric mean of perplexity over the three model sizes. The red line denotes aggregate ΔLogLik from GPT-2 Small surprisal (Radford et al., 2019) for reference. See Appendix A for the results on each individual dataset.

generally replicate the peak observed prior to LM training, this peak becomes less pronounced with the *Medium* and *Large* LMs. In particular, the *Large* LMs show much smaller differences in both perplexity and ΔLogLik across different vocabulary sizes. This suggests that increased model sizes allow LMs to learn qualitatively similar predictions that overcome the initial biases imposed by token granularity. However, considering the average over model sizes (i.e. the purple points in Figure 2), we conclude that the token granularity represented by a vocabulary size of around 8,000 results in surprisal estimates that are the strongest predictors of naturalistic reading times, even over the widely-used GPT-2 Small (Radford et al., 2019).

4 Experiment 2: Impact on Magnitude of Surprisal-Based Garden-Path Effects

The second experiment evaluates the effect of token granularity on the magnitude of surprisal-based estimates of garden-path effects (GPE). The aim of this experiment is to evaluate how token granularity influences the LMs' sensitivity to syntax by evaluating their surprisal on more targeted syntactic constructions.

4.1 Procedures

We estimated surprisal-based GPE from the LMs trained in Section 3.2, using the data and following the modeling procedures of Huang et al. (2024).

Surprisal-RT Linking Function. First, to estimate a linking function between LM surprisal and human reading times, LME models were fit to raw SPR times (n=995, 814) of filler items (i.e. 'ordinary' sentences that do not incur processing difficulty due to syntactic disambiguation) drawn from the Provo Corpus (Luke and Christianson, 2018) that are provided by Huang et al. (2024). These filler LME models include LM surprisal and log frequency of the current word and two previous words, word length in characters, and index of word position within the sentence as main effects, as well as by-subject and by-item random intercepts. These modeling choices make similar assumptions about the functional form between surprisal and reading times and the lingering influence of previous words as Experiment 1.

Garden-Path Stimuli and Reading Time Data. The stimuli used in Huang et al. (2024) consist of 24 items of the Main Verb/Reduced Relative (MV/RR), Direct Object/Sentential Complement (NP/S), and Transitive/Intransitive (NP/Z) garden-path constructions. Each item consists of a sentence in the ambiguous condition and a sentence in the unambiguous control condition (Table 4). These sentences were read by a total 2,000 subjects

Construction/Condition	Example
MV/RR Ambiguous	The suspect sent the file *deserved further investigation* given the new evidence.
MV/RR Unambiguous	The suspect who was sent the file *deserved further investigation* given the new evidence.
NP/S Ambiguous	The suspect showed the file *deserved further investigation* during the murder trial.
NP/S Unambiguous	The suspect showed that the file *deserved further investigation* during the murder trial.
NP/Z Ambiguous	Because the suspect changed the file *deserved further investigation* during the jury discussions.
NP/Z Unambiguous	Because the suspect changed, the file *deserved further investigation* during the jury discussions.

Table 4: Examples of garden-path constructions studied in Huang et al. (2024). In each sentence pair, the critical word is highlighted in magenta, and its two spillover words are italicized. In the ambiguous conditions, the critical word disambiguates the syntactic structure of the sentence and incurs processing difficulty.

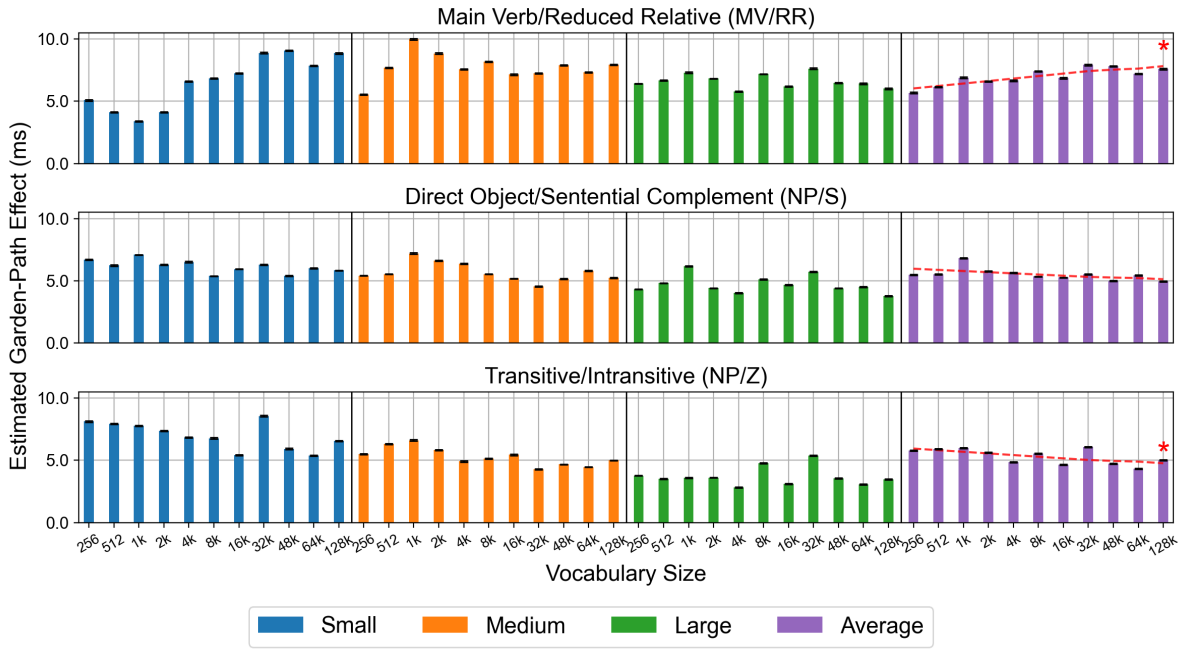

Figure 3: Estimated garden-path effects at the first spillover word (*further* in Table 4) for the three garden-path constructions using LM surprisal. Error bars denote 95% confidence intervals. The red curve shows predictions from the line of best fit with $\log_2(|V|)$ as the independent variable. $*$: $p < 0.05$.

using the SPR paradigm, which resulted in 47,695, 47,699, and 47,711 data points for the disambiguating critical word and its two spillover words respectively after data preprocessing.

Estimation of Surprisal-Based GPE. The LME models fit to SPR times of filler items were then used to generate predicted reading times (in ms) for sentences in the ambiguous condition and the unambiguous control condition. Subsequently, the increase in the predicted reading times due to the increase in surprisal across conditions at the critical word and two spillover words was estimated as the magnitude of surprisal-based GPE. To this end, another set of LME models that include a binary ambiguity condition, along with by-subject and by-item random intercepts was fit to the predicted reading times at each word for each construction.[9]

4.2 Results

The GPE estimated at the first spillover word[10] in Figure 3 again reveals an interaction between model size and token granularity, although the general trend is a lot less clear compared to Experi-

[9]Unlike Huang et al. (2024), who estimated the GPE of all three constructions simultaneously through one regression model that used dummy coding for constructions, we fit separate LME models to each subset in order to not impose any dependence between each construction's estimated GPE. Additionally, the design of both the 'filler item' and 'increase in predicted reading times' LME models had to be simplified from the original specifications in Huang et al. (2024) due to convergence issues.

[10]Effects are reported at the first spillover word as this region is where humans demonstrate the strongest GPE. See Appendix B for the results at the critical word and the second spillover word.

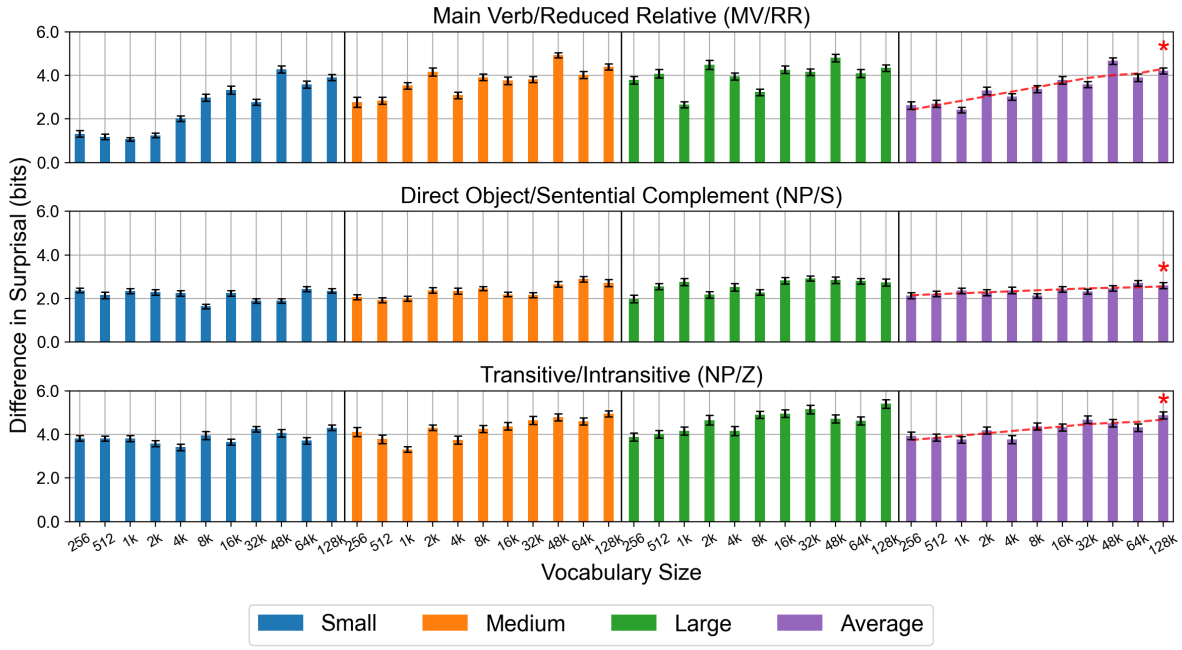

Figure 4: Increase in LM surprisal of the critical word across conditions for the three garden-path constructions. Error bars denote 95% confidence intervals. The red curve shows predictions from the line of best fit with $\log_2(|V|)$ as the independent variable. $*$: $p < 0.05$.

ment 1. The most notable trend is that the *Small* LMs trained with larger vocabulary sizes demonstrate larger GPEs on the MV/RR condition compared to their counterparts trained with smaller vocabulary sizes. This is consistent with the idea that having LMs treat words as independent symbols and learn from co-occurrences between them results in stronger representations of syntax. In contrast, the *Medium* LM trained with a vocabulary size of 1,000 appears to represent a peak in GPE across the three constructions, which demonstrates the opposite trend.

However, the raw difference in surprisal across conditions visualized in Figure 4 suggests that this peak is rather due to the difference in the estimated surprisal-reading time linking functions (i.e. coefficients and spillover dynamics of various predictors); *Large*, *Medium*, and *Small* LMs trained with larger vocabulary sizes generally show larger differences in surprisal at the critical word across all three conditions.

Moreover, consistently with Experiment 1, the trend in estimated GPE or the difference in surprisal across conditions is the least clear in *Large* LMs, although the models trained with larger vocabulary sizes tend to show larger differences in surprisal at critical words of NP/Z constructions. Given sufficient model sizes, it may be the case that LMs learn predictions that are even more similar for short, isolated sentences like the garden-path stimuli compared to longer, more extended naturalistic corpora. Finally, the manipulation of token granularity does not seem to alleviate the neural LMs' underestimation of human-like garden-path effects, which have been shown to be one or two orders of magnitude higher in previous work (van Schijndel and Linzen, 2021; Arehalli et al., 2022; Huang et al., 2024).

5 Conclusion

The influence of subword token granularity over which LMs are trained has been overlooked in cognitive modeling. Nonetheless, this granularity directly encodes statistical information about word length and frequency into word probabilities. Additionally, the granularity of tokens determines the collocational statistics between tokens within training corpora and ultimately impacts the quality of vector representations that are learned by LMs.

This work examines the influence of token granularity on the predictive power of LM surprisal on both naturalistic corpora and garden-path stimuli. Experiments with naturalistic reading times reveal a substantial influence of token granularity both prior to and after LM training. Tokens that are more fine-grained than contemporary standards

resulted in LM surprisal that is most predictive, which suggests that the information about word length and frequency encoded by the tokenization process correlates with processing difficulty.

In contrast, LMs trained on more coarse-grained tokens generally assigned higher surprisal to critical regions of garden-path constructions. This may be due to the more direct word-to-word associations learned by LMs, which is facilitated by tokens that are more word-like. As the critical word is identical across conditions in the garden-path stimuli, word length and frequency information appear to matter less in accounting for GPE. Taken together, these results suggest a large role of token granularity on LM surprisal for cognitive modeling.

Acknowledgments

We thank the ARR reviewers and the area chair for their helpful comments. This work was supported by the National Science Foundation (NSF) grant #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the NSF. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

Limitations

The influence of token granularity on the predictive power of surprisal identified in this work is supported by experiments using language models trained on English text and data from human subjects that are native speakers of English. Therefore, it remains to be seen whether the findings will generalize to language models and data collected in other languages. Additionally, although language models of multiple sizes were trained and evaluated in this work, models that are smaller or larger may yield different conclusions about the role of token granularity. Finally, this work is concerned with the use of language models as cognitive models of human sentence processing, and therefore does not relate to their use in natural language processing applications.

Ethics Statement

This work used data collected as part of previously published research (Futrell et al., 2021; Smith and Levy, 2013; Cop et al., 2017; Kennedy et al., 2003; Luke and Christianson, 2018; Huang et al., 2024). Readers are referred to the respective publications for more information on the data collection and validation procedures. As this work focuses on studying the connection between conditional probabilities of language models and human sentence processing, its potential negative impacts on society appear to be minimal.

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*, pages 301–313.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623.

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*, 68:255–278.

Jason J. S. Barton, Hashim M. Hanif, Laura Eklinder Björnstöm, and Charlotte Hills. 2014. The word-length effect in reading: A review. *Cognitive Neuropsychology*, 31(5–6):378–412.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 2397–2430.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624.

Agatha Christie. 1920. *The Mysterious Affair at Styles*. John Lane. Retrieved from Project Gutenberg.

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.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 10041–10071.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. On the proper treatment of tokenization in psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18556–18572.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

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.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 1–8.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Marcel Adam Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

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

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*, pages 66–75.

Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Sander Land and Max Bartolo. 2024. Fishing for Magikarp: Automatically detecting under-trained tokens in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11631–11646.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11251–11260.

Byung-Doh Oh, Christian Clark, and William Schuler. 2021. Surprisal estimators for human reading times need character models. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3746–3757.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921.

Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Keith Rayner, Marcia Carlson, and Lyn Frazier. 1983. The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, 22(3):358–374.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Cory Shain. 2024. Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, 8:177–201.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Tim Vieira, Ben LeBrun, Mario Giulianelli, Juan Luis Gastaldi, Brian DuSell, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. 2024. From language models over tokens to language models over characters. *arXiv preprint*, arXiv:2412.03719.

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*, pages 1707–1713.

Ethan Gotlieb Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023a. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511.

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

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 5184–5207.

A By-Corpus Regression Results

ΔLogLik for each LM surprisal evaluated in Experiment 1 on each reading time dataset is presented in Table 5.

B GPEs Estimated at Other Word Regions

The GPEs estimated at the critical word and the second spillover word are visualized in Figures 5 and 6 and respectively.

Model / $\lvert V\rvert$	NS	Brown	GECO_{FP}	GECO_{GP}	$\text{Dundee}_{\text{SP}}$	$\text{Dundee}_{\text{FP}}$	$\text{Dundee}_{\text{GP}}$	Provo_{SP}	Provo_{FP}	Provo_{GP}	Total
Small 256	1558.0	473.5	495.2	157.4	**137.4**	**401.9**	170.5	4.3	212.6	64.2	3675.0
Small 512	1594.0	489.2	485.4	156.5	127.4	399.4	166.3	5.0	214.7	76.1	3714.0
Small 1000	1666.0	498.5	534.6	171.8	127.7	399.8	178.1	4.3	222.7	72.0	3875.5
Small 2000	1650.0	484.7	535.6	164.6	132.0	376.4	165.2	4.2	230.2	75.4	3818.3
Small 4000	**1681.0**	484.7	551.5	170.1	128.4	378.6	168.0	5.3	211.3	75.3	3854.2
Small 8000	1668.0	**499.1**	**598.5**	188.2	134.0	400.6	177.1	7.9	219.4	84.5	**3977.3**
Small 16000	1599.0	480.3	552.8	169.4	136.3	399.4	**181.6**	**10.0**	**251.5**	89.3	3869.6
Small 32000	1561.0	465.8	538.3	190.2	124.7	377.9	172.7	**10.0**	231.9	**100.0**	3772.5
Small 48000	1510.0	462.1	550.5	**193.3**	126.4	359.6	160.0	8.3	210.6	92.8	3673.6
Small 64000	1497.0	450.6	548.4	173.8	112.1	361.8	158.8	6.3	191.3	82.2	3582.3
Small 128000	1556.0	440.6	523.6	181.0	116.4	345.9	151.8	9.7	185.1	86.1	3596.2
Medium 256	1582.0	470.8	538.5	180.4	124.6	361.8	152.9	9.8	192.3	80.0	3693.1
Medium 512	1613.0	485.1	534.9	177.6	129.6	347.4	153.8	6.2	205.3	78.7	3731.6
Medium 1000	**1626.0**	**504.4**	522.1	186.6	121.8	350.6	156.0	6.0	203.4	82.2	**3759.1**
Medium 2000	1541.0	485.6	549.9	183.9	115.7	342.1	157.0	9.4	206.0	94.4	3685.0
Medium 4000	1587.0	471.3	537.6	170.9	117.8	342.0	157.4	11.9	202.1	95.4	3693.4
Medium 8000	1579.0	496.2	**564.3**	185.1	125.3	344.1	153.3	11.2	**207.4**	83.6	3749.5
Medium 16000	1454.0	489.6	521.7	184.8	126.7	349.1	161.0	11.0	174.4	88.2	3560.5
Medium 32000	1529.0	479.3	523.3	200.9	**131.0**	**370.3**	**163.7**	10.0	202.9	**96.1**	3706.5
Medium 48000	1472.0	473.3	547.5	194.4	129.6	359.3	156.6	9.8	169.6	77.0	3589.1
Medium 64000	1447.0	463.4	508.6	189.2	127.7	344.2	153.5	9.3	172.1	83.1	3498.1
Medium 128000	1440.0	470.1	488.5	**201.5**	127.3	336.8	153.6	**12.3**	178.7	90.3	3499.1
Large 256	1553.0	473.0	**534.4**	184.0	120.6	326.6	147.3	9.4	175.0	75.7	3599.0
Large 512	1531.0	482.5	503.1	179.5	**136.3**	**336.9**	153.9	7.6	177.3	83.1	3591.2
Large 1000	**1603.0**	479.5	519.4	188.2	121.1	323.7	152.1	6.8	168.6	80.8	**3643.2**
Large 2000	1558.0	479.6	513.4	165.5	122.7	326.3	147.5	8.5	172.9	77.3	3571.7
Large 4000	1549.0	465.7	506.6	185.6	119.0	325.5	149.2	8.7	163.9	82.9	3556.1
Large 8000	1536.0	**490.5**	511.5	185.4	119.8	335.3	153.9	10.6	178.7	85.2	3606.9
Large 16000	1426.0	473.7	494.9	**201.2**	132.5	333.4	**159.1**	10.7	**179.8**	**89.8**	3501.1
Large 32000	1428.0	487.8	498.5	190.0	124.4	332.5	155.7	9.2	176.9	84.0	3487.0
Large 48000	1446.0	467.0	483.0	193.6	124.1	332.5	157.6	**11.2**	157.1	83.0	3455.1
Large 64000	1410.0	466.3	491.2	188.2	124.1	332.9	150.7	8.1	146.5	82.6	3400.6
Large 128000	1381.0	462.8	480.4	199.3	133.7	324.8	155.0	8.7	151.1	81.7	3378.5
GPT-2 Small	1459.0	543.8	463.0	209.0	151.6	343.8	181.3	10.3	224.5	113.7	3700.0

Table 5: ΔLogLik of each LM surprisal evaluated in Experiment 1 on the exploratory partition of each reading time dataset. NS: Natural Stories.

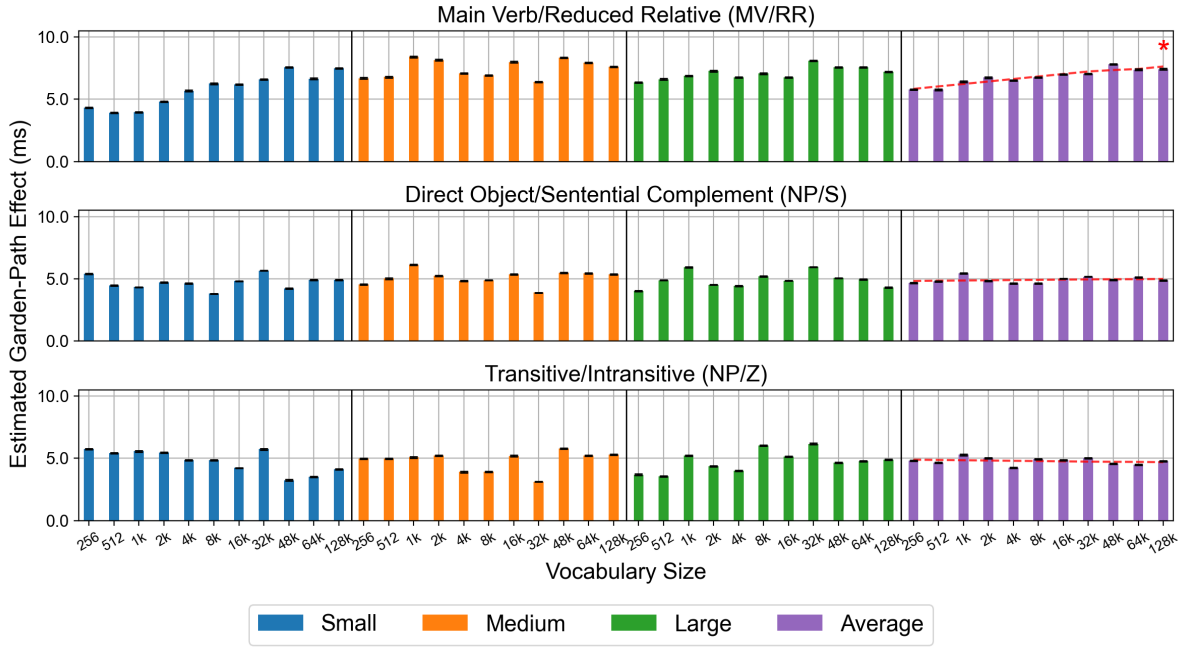

Figure 5: Estimated garden-path effects at the critical word (*deserved* in Table 4) for the three garden-path constructions using LM surprisal. Error bars denote 95% confidence intervals. The red curve shows predictions from the line of best fit with $\log_2(|V|)$ as the independent variable. $*$: $p < 0.05$.

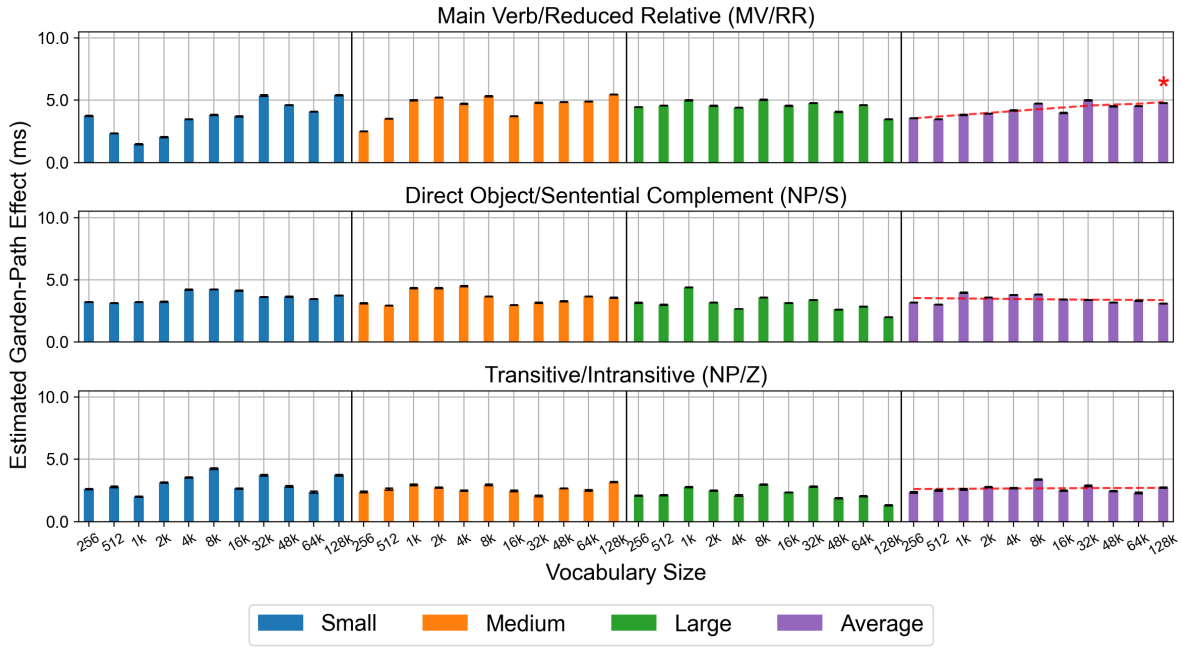

Figure 6: Estimated garden-path effects at the second spillover word (*investigation* in Table 4) for the three garden-path constructions using LM surprisal. Error bars denote 95% confidence intervals. The red curve shows predictions from the line of best fit with $\log_2(|V|)$ as the independent variable. $*$: $p < 0.05$.