Word predictability estimates from language models are not robust to tokenizer vocabulary

Kien Nguyen
Harvard University
kientnguyen.ntk@dgmail.com

Suhas Arehalli
Macalester College
sarehall@macalester.edu

Abstract

Much recent work has been interested in modeling language processing using measures of predictability estimated from pretrained language models. These models, however, are primarily built as language technologies rather than cognitive models, and make many design choices that may align poorly with theories of human language processing. We investigate one such choice — the size of the vocabulary learned by a BPE tokenizer — and investigate (1) its effect on the linguistic plausibility of subword units the model learns, (2) whether vocabulary size has a substantial influence on the surprisal estimates a model generates, and (3) whether those differences in surprisal translate to differences in the quality of downstream reading time predictions. We find that while vocabulary size doesn't substantially affect the rate of morphologically reasonable tokenizations, it does have an impact on surprisal estimates and reading time predictions from 5-gram, LSTM, and GPT-2 language models. Moreover, we find that these differences primarily affect words that are split by the tokenizer, suggesting that psycholinguists should take care to design stimuli meant for computational modeling with subword tokenization in mind.

1 Introduction

Within psycholinguistics, there has been a surge of interest in using language models to predict aspects of human language processing (Wilcox et al., 2023; Shain et al., 2024; Huang et al., 2024; Wilcox et al., 2024; Kuribayashi et al., 2025, etc.). This work typically involves deriving some measure of each word's probability in context (surprisal, entropy, etc.) from a neural language model and then using that measure to predict some measure of human processing difficulty (reading times, rates of regressive eye movements, or accuracy, among others). While many of these measures are motivated by psycholinguistic theory (i.e., Surprisal Theory, Levy, 2008), the process of deriving these measures often involves steps that are a product of model training conventions that are motivated by their ability to facilitate performance rather than their suitability for cognitive modeling.

One such step in the model training pipeline is *tokenization*, which segments raw text input into atomic units for further processing. Modern language models often use subword tokenization strategies like BPE (Sennrich et al., 2016), which constructs a vocabulary of possible units based on statistical co-occurrences between characters in a training corpus. For a computational psycholinguist, these schemes may seem appealing: Building units based on co-occurrences bears some resemblance to the role of transitional probabilities in statistical learning theories of language acquisition (Saffran et al., 1996; Saffran, 2001), and the use of subword units allows for models to capture morphological information shared between words.

In practice, however, there are many unappealing aspects to BPE tokenization. For instance, BPE is known to construct subword tokenization splits that often do not reflect linguistically sensible decompositions. Additionally, while Nair and Resnik (2023) found that subword tokenization choices had little impact on the quality of predictions on broad-coverage psycholinguistic data, they found that this result was driven by the fact that standard BPE tokenizers only rarely split words into multiple tokens. This lack of subword splits generated by standard practice leads to two potential issues for psycholinguistic modeling: First, this limits the purported benefits of shared subword information across words, weakening the ability of models to account for subword processing mechanisms human speakers may use (Fiorentino and Poeppel, 2007; Ettinger et al., 2014; Bunzeck and Zarrieß, 2025, etc.). Second, this leads to a reasonable worry that the few words that are actually split will be modeled poorly, as their scarcity would give

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 34–44

little incentive for models to account for differences based on subword splits during fitting (as Nair and Resnik, 2023 observed).

This work aims to explore the relationship between the number of subword splits a tokenizer creates, the morphological sensibility (or lack thereof) of the splits, and the ability of models trained with these tokenizers to model human reading times. In particular, we utilized the *vocabulary size* parameter of BPE tokenizers to manipulate the number of subword splits, evaluating these tokenizers both on their ability to generate morphologically sensible splits, the robustness of their surprisal estimates, and whether surprisals of split and unsplit words share the same relationship with human reading times.

In the first experiment, we quantify the exact effect of vocabulary size on the number of subword splits and their morphological sensibility by training 100 different BPE tokenizers, ranging from the minimum vocab size of 257 to 50257, the size used by the GPT-2 LMs (Radford et al., 2019). We find that as vocabulary size increases, the number of words being split decreases rapidly, but the ratio of morphological splits stabilizes quickly. To further investigate how vocabulary sizes affect language modeling, in the second experiment, we trained 5-gram, LSTM, and GPT-2 models on four representative tokenizer sizes. We find that the word-by-word surprisals generated by these models were not robust to the tokenizer's vocabulary size. In the final experiment, we attempt to predict reading times using our LSTM models' surprisals under the assumptions of surprisal theory, aiming to evaluate whether the conversion factor between split word surprisal and reading times is the same as that between unsplit words and reading times. We find that split words' reading times are consistently under- or overpredicted relative to unsplit words. Taken together, our results suggest that choices in tokenizer implementation — in particular, the vocabulary size used — can impact model surprisal estimates and the subsequent quality of reading time predictions in psycholinguistics modeling. Moreover, we advise experimenters to be particularly careful when doing targeted evaluations on a small number of experimental sentences, as the estimates of infrequent split words that may appear in these materials may be most impacted by tokenization choices.[1]

[1]Code to replicate our analyses is available here.

2 Experiment 1: How morphological is BPE tokenization at lower vocabulary sizes?

Nair and Resnik (2023) found that the non-morphological nature of BPE subword tokenization does not appear to affect the broad-coverage predictive power of models of reading times. However, they also note that due to their use of a standard, pretrained tokenizer, very few words in the corpus they tokenized were actually split. As a result, their broad-coverage analyses of subword tokenization are largely an evaluation of word-level, rather than subword, tokenization.

Motivated by this, we aim to evaluate whether tokenizers that split words to a greater degree may differ in both their effect of psycholinguistic modeling pipelines and their morphological validity. In this first experiment, we explore how manipulating the vocabulary size parameter of a BPE tokenizer could allow for more subword tokenization to take place in practice on the particular corpora we plan to analyze. We ask (1) to what extent are words actually tokenized into subword tokens at each vocabulary size and (2) to what extent is that subword tokenization morphologically sensible.

2.1 Methodology

Given our desire to understand the effects of tokenization choices on psycholinguistic experimentation, we evaluate our models using a pipeline designed to simulate how a tokenizer (and, in later experiments, a language model) might be used for psycholinguistic modeling. In particular, we simulate a procedure where language model is trained from scratch on a developmentally plausible dataset and is subsequently used to predict reading times from a broad-coverage, naturalistic corpus.

In this experiment, we thus train 100 BPE tokenizers on the BabyLM Corpus (Warstadt et al., 2023), a developmentally plausible corpus of 100 million words designed for psycholinguistic modeling, with vocabulary sizes ranging from 257 to 50257 (the vocabulary size used in GPT-2's tokenizer) in increments of 500. We then used these tokenizers to tokenize the Natural Stories Corpus (NSC; Futrell et al., 2017), a corpus of word-by-word readings times for a set of naturalistic, syntactically rich texts.

To evaluate the quality of the tokenization, we label each word in terms of the morphological sensibility of its tokenization using the umLabeller

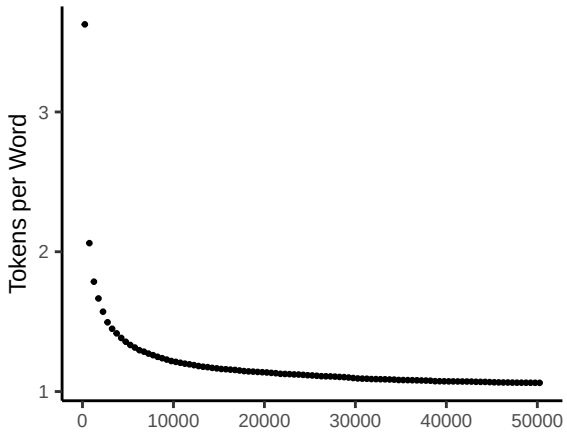

Figure 1: The ratio of tokens to words in the NSC at each BPE vocabulary size.

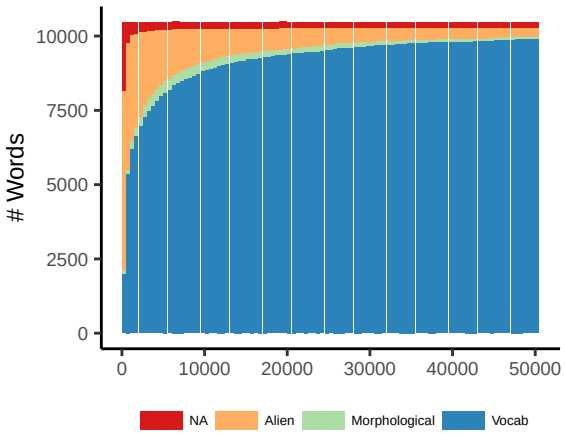

Figure 2: The number of words in the NSC with each label at each vocabulary size.

(Batsuren et al., 2024). Words were labeled *Vocab* if the word is unsplit at this vocabulary size. If the word is split, it is labeled *Morphological* if the split is consistent with a standard linguistic analysis and *Alien* otherwise. The label *NA* is reserved for words not in the labeller's vocabulary.

2.2 Results

We first consider the extent to which words are split at each vocabulary size. We plot the ratio of tokens to words at each vocabulary size in Figure 1. We find that the token-per-word ratio decreases rapidly as vocabulary size increases before beginning to plateau, starting at around a vocabulary size of 5000. Like Nair and Resnik (2023), we find that in large vocabulary sizes, the vast majority of words are not split, resulting in these BPE tokenizers effectively functioning as word-level tokenizers (Figure 2).

We then investigate the extent to which the subword tokenization we see at lower vocabulary sizes is morphological. We find that, regardless of vocabulary size, most of the splits are *Alien*, or non-morphological (i.e., Figure 2). When considering just the percentage of *Morphological* and *Alien* splits (Figure 3), we find that past a vocab size of 5000, while the raw number of split tokens continues to shrink, the proportion of those splits that are considered morphological remains stable.

3 Experiment 2: Does vocabulary size impact surprisal estimates?

Experiment 1 showed that the number of split words decreases as vocabulary size increases, but that, above a certain vocabulary size, the morphological sensibility of these splits is relatively stable. In particular, consistent with Nair and Resnik (2023), we find that a BPE tokenizer with a vocabulary size of 50257 (that of GPT-2), would functionally operate as a word-level tokenizer over a dataset like the NSC. In addition, we confirm that tokenizers with much smaller vocabulary sizes can provide a larger number of splits with roughly the same rate of morphological validity as larger vocabulary sizes.

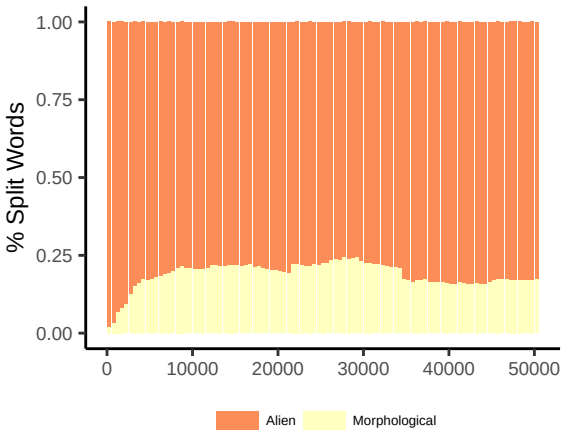

Figure 3: The percentage of each label among split words in the NSC, excluding those labeled *NA*.

Now that we have established that we can manipulate properties of psycholinguistic interest — how much subword information can be represented, and whether the subword units correspond to linguistically motivated decompositions — our attention turns to whether these properties of tokenizations actually affect the behavior of language models trained over them. In particular, we ask whether surprisals are robust to variation in vocabulary size,

with a particular focus on the extent to which small vocabulary sizes with greater levels of tokenization may lead to models that generate surprisals different from those produced by models trained using the 50257 vocabulary size used by GPT-2's tokenizer.

Crucially, a word's surprisal should be unaffected by tokenization choices[2] as the sum of all of the subword tokens' surprisals should be exactly the surprisal of the word. Therefore, differences in surprisal estimates for a fixed word in context across vocabulary sizes indicates a sensitivity to tokenization choices in the models we use to estimate surprisal.

3.1 Methodology

We select 4 representative tokenizers from Experiment 1 with which to tokenize our training data: those with vocabulary sizes 4257, 8257, 20257, and 50257. These, in increasing order, represent the size with the largest number of morphological tokens, the size with the largest morphological to alien token ratio, a small size that already has a small number of split tokens and morphological tokens, and the size used by GPT-2 tokenizer. We consider size 50257 as a baseline, as it is the standard vocabulary size of the tokenizer used by GPT-2, a widely used pretrained model. For each of these vocabulary sizes, we train both a 5-gram language model using KenLM (Heafield, 2011) with modified Kneser-Ney smoothing (Chen and Goodman, 1996) to match the analysis done in Nair and Resnik (2023), an LSTM language model with the architecture and hyperparameters of Gulordava et al. (2018), a neural LM also widely used in psycholinguistics work, and a transformer model with the architecture of GPT-2 Small.

We train the models on the four tokenizations of the BabyLM corpus. We then computed word-by-word surprisals over the NSC, summing the surprisals of subword pieces. As in Experiment 1, we label each word as morphologically split (*morph*), non-morphologically split (*alien*), or unsplit (*vocab*) using the umLabeller. As the results from the NSC for this and the following experiment emerged from an exploratory analysis of the relationship between our tokenizers' vocabularies and the resulting surprisals, we chose to run the same analyses of the LSTM models on the monolingual portion of GECO (Cop et al., 2017), a corpus of eyetracking data. We do this to evaluate the generalizability of our findings to a different set of materials and, in the following experiment, with different reading measures collected with a different methodology.

3.2 Results

Figure 4 shows the changes in surprisal (or Δsurprisal) for each word at a particular vocabulary size relative to the baseline vocabulary size of 50257, for all three models. We further subdivide based on whether the word is split at the particular vocabulary size considered and at the baseline: If a word's status (i.e. *split* or *unsplit*) is the same between the two sizes, they are categorized as either *Stay Unsplit* and *Stay Split*. Otherwise, they are categorized by which label they have at the smaller vocabulary size (i.e., before being merged), giving us the two additional categories *MorphToVocab*, *AlienToVocab*. To aid in readability, we do not plot words who are labeled *NA* by the umLabeller. This Δsurprisal measure allows us to see the extent to which moving from a standard large vocabulary size like 50k to a smaller vocabulary size would affect surprisal estimates for various types of words (i.e., those that would become split by virtue of swapping from a smaller to a larger vocabulary size). Note again that if surprisal estimates from the model were entirely robust to vocabulary size, then we should see Δsurprisals near 0.

For all models, the category that remains most consistent across vocabulary sizes is *Stay Unsplit*, with an average Δsurprisal nearest to 0. Notably, for LSTM models, *Stay Split* words are, like *MorphToVocab* and *AlienToVocab*, overestimated at smaller vocabulary sizes while they are underestimated by both 5-gram and GPT-2 models. We find no differences between the *MorphToVocab* and *AlienToVocab* words, which both overestimate surprisals relative to the baseline at all vocab sizes. Similarly, for LSTM models, while *Stay Unsplit* words are overestimated less than the other categories, all categories consistently overestimate surprisal compared to the baseline across vocabulary sizes. These results roughly hold for the GECO dataset, as shown in Figure 5, with the exception of the *Stay Split* category, which for size 20257 is not significantly different than baseline.

Taken together, these analyses indicate that manipulating vocabulary size has a consistent effect on the surprisals generated from a language model,

[2]Though note that there is some variation in how to quantify word-level surprisal from sub-word parts; for example, on how to handle whitespace (Oh and Schuler, 2024; Pimentel and Meister, 2024; Giulianelli et al., 2024).

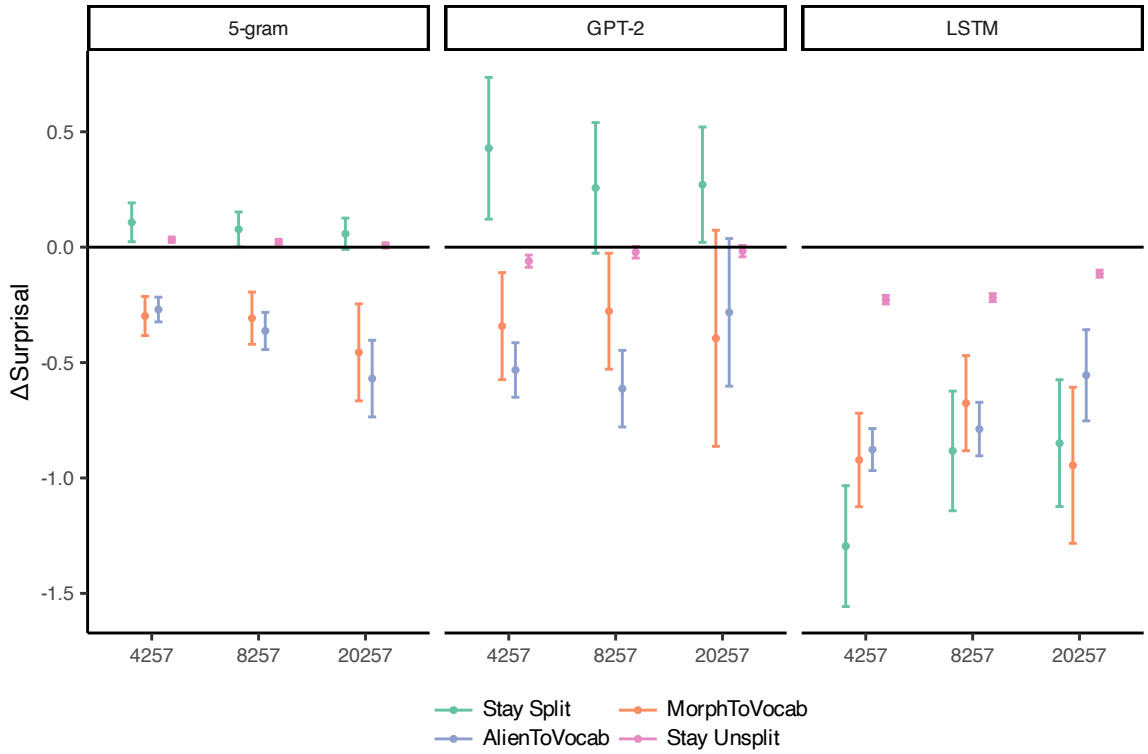

Figure 4: Δsurprisal of NSC across LSTM and 5-gram models by change in label. Error bars are 95% confidence intervals.

particularly for words that become a single token at the large vocabulary size. More specifically, we find that surprisals of split words at smaller vocabulary sizes tend to be larger than that of those same words unsplit at the larger 50257 GPT-2 vocabulary size. While these results are consistent with better language modeling performance (i.e., reduced perplexity) at higher vocabulary sizes, note that improvement in perplexity is not always tied to improvements in psychometric predictive power (Oh and Schuler, 2023), and this improvement in perplexity should not be taken as a sign of a better cognitive model.

4 Experiment 3: Does vocabulary size affect the ability of LMs to capture human reading behavior?

In Experiment 2, we found that manipulating vocabulary size for a language model's tokenizer has an impact on the surprisals that a language model generates. Moreover, these manipulations impacted words that were split at smaller vocabulary sizes more than words that were unsplit. While these results indicate that the *surprisals* produced by language models are not consistent at different vocabulary sizes, they do not lend any insight into whether some of these estimates were better than others, or whether these differences in surprisal result in qualitatively different reading time predictions.

In this experiment, we address these concerns by using the surprisals generated by these models to predict reading times under a surprisal-theoretic framework that assumes a fixed linear relationship between surprisals and reading times (Levy, 2008; Smith and Levy, 2013; van Schijndel and Linzen, 2021). In particular, motivated by the results of Experiment 2, we separately analyze the performance of these models on words that are split and those that are unsplit by the tokenizer at each vocabulary size to determine whether the model-estimated surprisals are comparable in their ability to predict the processing difficulty of words split and words unsplit by the tokenizer.

Critically, under this surprisal framework we would predict that surprisals for split and unsplit words have identical relationships with measures of processing difficulty. Thus, if a model consistently over- or underpredicted the reading times of split words we would have evidence that the surprisal estimates for the split words from that model were poor, as they were inconsistent with the estimates

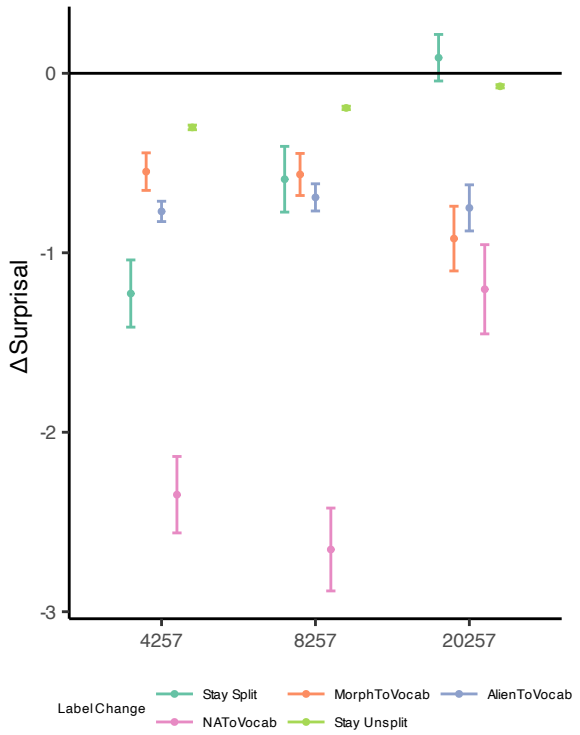

Figure 5: Δsurprisal of GECO across LSTM models by change in label. Error bars are 95% confidence intervals.

for unsplit words, and vice-versa.

4.1 Methodology

For each LM, we fit a linear regression model using surprisal to predict Reading Times from the NSC's Self-Paced Reading task, including log frequency and word length as additional predictors. As our goal is to look at the relationship between a word's status as split or unsplit by the tokenizer, its surprisal estimate, and the associated reading times from participants in the NSC. However, spillover effects in self paced reading result in some of the effect of a words surprisal being delayed to subsequent words, meaning that the standard surprisal analysis that acccounts for these spillover effects typically include the surprisals of the prior two or three words as predictors, which will often include both split and unsplit words. This would mean we would be unable to isolate the impact of a particular word (and its split or unsplit status) on reading times.

To circumvent this, we use a restricted model that predicts the current word's reading time using only a single surprisal taken from either the current word or one of the prior two words. We determine which word's surprisal to use by fitting an initial linear regression model that includes terms for the current and prior two word's surprisal, log-frequency, and length. The surprisal term with the largest coefficient is then chosen as the one whose word's label and surprisal is used in the final models.[3] The goal of this selection is to match each word's surprisal with the single reading time that most reflects the processing difficulty associated with processing it. Critically, we will evaluate the resulting model not on the strength of the relationship between surprisal and reading times, but on whether split or unsplit words are predicted equally well. For the NSC, we find that the previous word's surprisal was most predictive, and thus include it in the final models.

We then fit two linear models using that model formula, an UNSPLIT model fit on only unsplit words and a FULL model fit on the entire dataset. The FULL model is intended to capture behavior in our simulated psycholinguistic modeling experiment, looking at how the impact of each word's tokenization impacts the ability of surprisal to predict its reading time.

On the other hand, the UNSPLIT model is designed to evaluate how different the mapping from surprisal to RT is for split and unsplit words. If unsplit word's surprisals have a similar relationship with reading times as split word's surprisals (as one would assume under an analysis that does not consider tokenization) then a model estimated using only unsplit words should generalize to split words. On the other hand, if their surprisals have distinct relationships with reading times, then we should see consistent over- or underestimation of reading times.

Similar to Experiment 2, we validate our findings on the GECO eyetracking corpus, fitting new linear models using surprisal to predict Gaze Duration, Go Past Time and Total Reading Time from the monolingual dataset. As with NSC, we selected one surprisal term for each model to account for spillover: Previous word's surprisal was the best predictor for Go Past Time and that the current word's surprisal was the best predictor for Gaze Duration and Total Reading Time.

4.2 Results

Figure 6 shows the residual errors (true RT - predicted RT) of the UNSPLIT model and the FULL model predicting NSC Self-Paced Reading times

[3] with the final model formula for a word at position i being $RT_i \sim Surprisal_{best} + Length_i * Frequency_i + Length_{i-1} * Frequency_{i-1} + Length_{i-2} * Frequency_{i-2}$

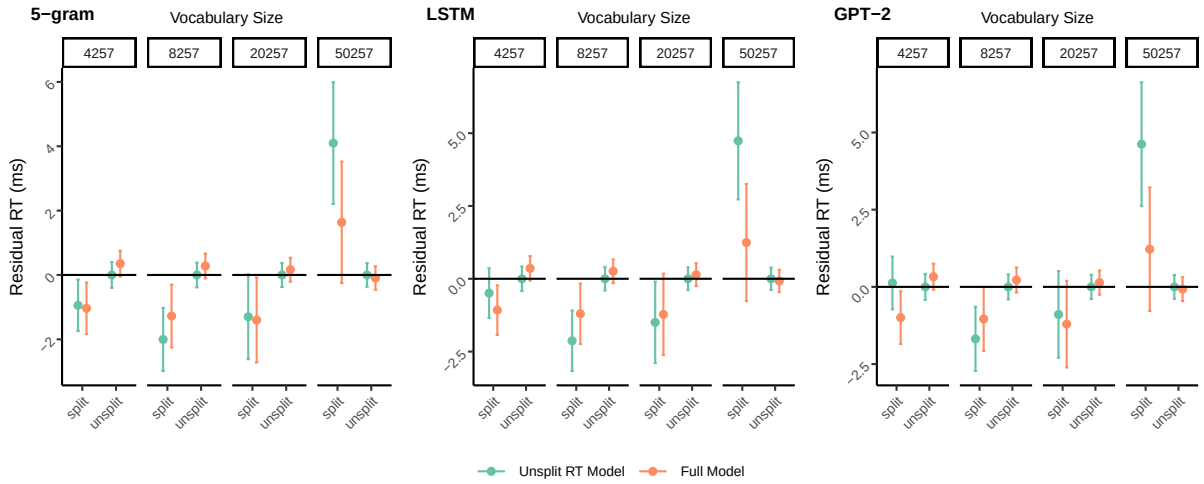

Figure 6: Residual error of FULL Model and UNSPLIT Model predicting NSC Reading Time across vocabulary size for 5-gram, LSTM, and GPT-2 models. Error bars are 95% confidence intervals.

across vocabulary sizes for 5-gram, LSTM, and GPT-2 models. For both the 5-gram and the LSTM models, the UNSPLIT model consistently overpredicts the split words across all vocabulary sizes except for the largest, 50257. For GPT-2, we see a similar trend, but with no significant under- or overprediction at vocabulary sizes 4257 and 20257. Unsplit words, on the other hand, show no bias in model errors.

These patterns of under and overprediction of split words are consistent with the idea that, for the UNSPLIT model, split words are unseen *out-of-distribution* data: The surprisal estimates for split words have a different relationship with reading times than unsplit words. Further, these patterns arise even in the FULL models that are trained on all words.

When considering the impact of vocabulary size, we see that at the three smaller vocabulary sizes, we see a consistent overprediction. At a vocabulary size of 50257, the pattern of overprediction reverses, with split words being largely *underpredicted*.

Since we analyze tokenization using BPE, words that are unsplit at a particular vocabulary size are necessarily more frequent in our training corpus than those that are split. Thus, these results are consistent with the finding that less *frequent* words are over or underpredicted under by our regression models. However, under typical surprisal-theoretic modeling assumptions (Levy, 2008; Smith and Levy, 2013; Shain et al., 2024, etc.), surprisal and reading times should have a fixed linear relationship regardless of frequency, and thus we should not see bias in infrequent words outside of those captured by the surprisal and frequency terms in our regression.

The residual errors of the UNSPLIT models and FULL models' predicting Gaze Duration, Go Past Time and Total Reading Time from GECO are shown in Figure 7.

Our observation of bias in model error on split words in NSC Self Paced Reading Time is consistent with our findings in GECO. In both cases, unsplit words are predicted with little bias in residual error in both unsplit and full models, while split models are consistently under- or overpredicted. However, we see great variability in patterns of under- and overprediction depending on both measure and paradigm: We see mostly *under*prediction in eyetracking measures, and see underprediction reducing with increasing vocabulary size in Gaze Duration, but increasing in Go Past time and for the FULL models in Total Reading Time. Regardless of this variation, these results are consistent with the claim that split words' surprisals have a different relationship with reading times than those of unsplit words under a surprisal-theoretic modeling framework.

5 Discussion

To summarize, across three experiments, we find that BPE vocabulary size influences the estimated surprisals of words and predicted reading times (from both self-paced reading and eyetracking data) under a surprisal-based modeling framework. In Experiment 2, we find that surprisal estimates for specific words differ across tokenization sizes, but

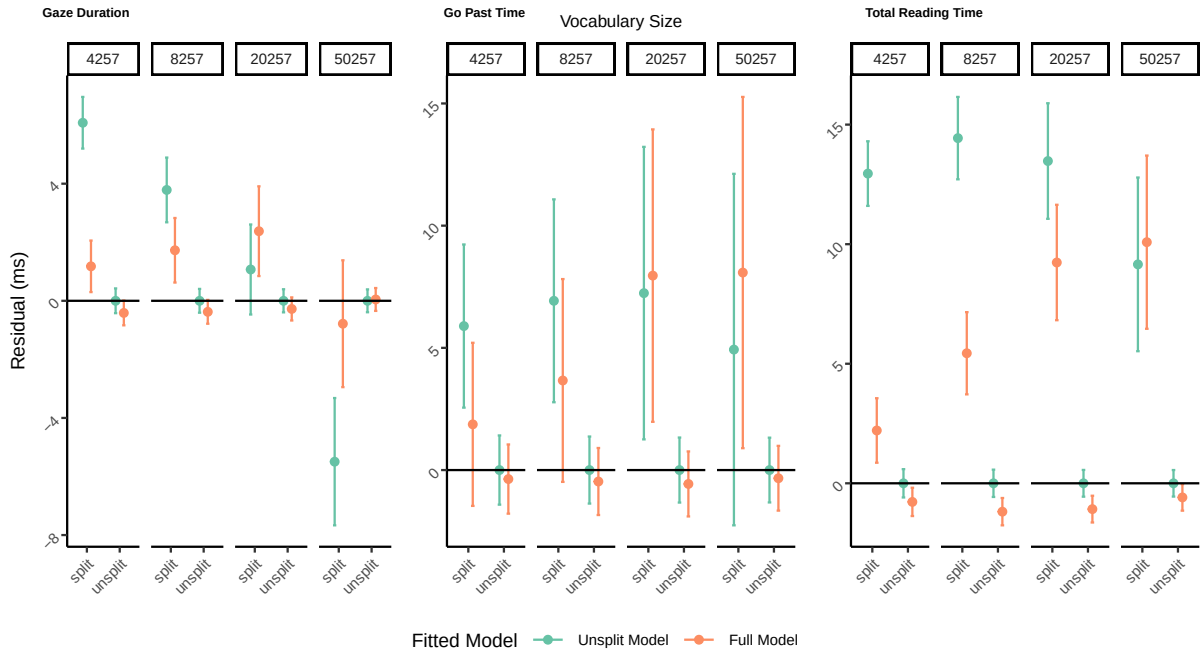

Figure 7: Residual error of FULL Model and UNSPLIT Model Predicting GECO Reading Time measures across vocabulary size for LSTM models. Error bars are 95% confidence intervals.

these differences are primarily driven by words that are split at smaller vocabulary sizes getting single-token representations at larger vocabulary sizes. In Experiment 3, we find that surprisal-based models of reading times consistently under- or overpredict reading times of split words relative to unsplit words.

These findings illustrate the importance of carefully considering the impact of technical parameters like tokenizer vocabulary size when using language models for psycholinguistic modeling; word-level surprisals and downstream measures of difficulty derived from an off-the-shelf model can only be understood with respect to its tokenization choices. We find that particular classes of words — here, words that may be either split or unsplit depending on the choice of vocabulary size — may be particularly impacted by these tokenization choices. Practically, this suggests that psycholinguists using models to generate surprisal estimates should consider designing stimuli that are robust to these choices by, for example, avoiding words that are split by the tokenizer.

Our results expand on and complement prior work investigating the role of tokenization on psycholinguistic modeling. Nair and Resnik (2023) analyzed the impact of various tokenization algorithms (orthographic/character-level, BPE, and morphological) using 5-gram models and, consistent with our results, found that off-the-shelf BPE tokenizers were worse predictors of split word reading times than unsplit words. Most similar to this work, Oh and Schuler (2025) provided both a broad coverage evaluation of the ability of a suite of State-Space Models (SSMs) with various vocabulary sizes to predict reading time data and a targeted case study on using these models to predict experimental data on Garden Path sentences. Similar to this work, they found a substantial impact of vocabulary size, and find the best fit to human reading times involves vocabulary sizes substantially smaller than that of off-the-shelf models. Among this landscape, our work provides a few unique contributions: First, we provide converging evidence from LSTM models, which are both popular and well understood in psycholinguistic modeling (Gulordava et al., 2018; Chowdhury and Zamparelli, 2018; Marvin and Linzen, 2018; Wilcox et al., 2018; Lakretz et al., 2019; Futrell et al., 2019; van Schijndel and Linzen, 2021; Wilcox et al., 2024, etc.) and Transformer models using the GPT-2 small architecture. Second, we provide analyses targeted at tracking the way particular word's surprisal estimates and reading time predictions are impacted by vocabulary choices, which complements the more broad-coverage analyses done by Oh and Schuler (2025) and Nair and Resnik (2023).

These results, taken together, underscore the

need for a more thorough understanding of the role tokenizers play in predicting human reading times with language models. Most notably, following Nair and Resnik (2023) and Beinborn and Pinter (2023), we hope to see more work understanding the extent to which tokenization and morphological processing interact, particularly the influence of morphological tokenization on the modeling of phenomena like subject-verb agreement, which has been a target of much surprisal-based modeling (Linzen et al., 2016; Gulordava et al., 2018; Lakretz et al., 2019; Finlayson et al., 2021; Arehalli and Linzen, 2024, etc.). Additionally, we see promise in the work of developing efficient and broad-coverage tokenization methods that better capture the morphological assumptions made in psycholinguistic work to provide an alternative to off-the-shelf models developed without cognitive plausibility as a goal.

Limitations

This work analyzes a small set of models of three architectures on a pair of corpora collected using two behavioral methodologies and tokenized using BPE and thus may not be representative of broader classes of models and tokenization algorithms that may be used to model a wider set of stimuli. We hope that future work, in concert with existing results from others, can help establish the robustness of these findings. Similarly, these experiments were only conducted on English text, and these findings may not translate to languages with different morphological or orthographic systems.

References

Suhas Arehalli and Tal Linzen. 2024. Neural Networks as Cognitive Models of the Processing of Syntactic Constraints. *Open Mind*, 8:558–614.

Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. 2024. Evaluating subword tokenization: Alien subword composition and OOV generalization challenge. *Preprint*, arXiv:2404.13292.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478–4486, Singapore. Association for Computational Linguistics.

Bastian Bunzeck and Sina Zarrieß. 2025. Subword models struggle with word learning, but surprisal hides it. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 286–300, Vienna, Austria. Association for Computational Linguistics.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, California, USA. Association for Computational Linguistics.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Allyson Ettinger, Tal Linzen, and Alec Marantz. 2014. The role of morphology in phoneme prediction: Evidence from MEG. *Brain and Language*, 129:14–23.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. 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 1828–1843, Online. Association for Computational Linguistics.

Robert Fiorentino and David Poeppel. 2007. Compound words and structure in the lexicon. *Language and Cognitive Processes*, 22(7):953–1000.

Richard Futrell, Edward Gibson, Hal Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2017. The natural stories corpus. *Preprint*, arXiv:1708.05763.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Miami, Florida, USA. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *Preprint*, arXiv:1803.11138.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

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.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large Language Models Are Human-Like Internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

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

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

Byung-Doh Oh and William Schuler. 2025. The impact of token granularity on the predictive power of language model surprisal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4150–4162, Vienna, Austria. Association for Computational Linguistics.

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

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

Jenny R Saffran. 2001. Words in a sea of sounds: The output of infant statistical learning. *Cognition*, 81(2):149–169.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

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

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.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 55(4):805–848.