Words, Subwords, and Morphemes: Surprisal Theory and Units of Prediction

Sathvik Nair[1], Colin Phillips[2,1], and Philip Resnik[1] (University of Maryland[1] & Oxford University[2])

Surprisal theory, which relates processing difficulty of linguistic information with its probability given context, has been useful to quantify prediction in human sentence processing [11]. Earlier work used probabilities estimated from statistical *n*-gram models, which take whole words as input [10]. However, current methods of computing surprisal are estimated from neural network-based language models (NLMs) which do not assign probabilities to words, but to subword tokens [7, 9, 13] based on frequencies of character sequences, which may or may not correspond to linguistically meaningful units (cf. [4]). For instance, *decomposition* could be broken into *dec*, *om*, and *position* instead of its constituent morphemes. If a word is split into multiple tokens, many approaches sum the surprisals of the individual tokens [6, 11], in line with assuming that cognitive effort dedicated to a word is proportional to the sum of the effort on its parts [8]. There is evidence human comprehension involves prediction [5] and composition [8] at the morphological level, but if subword tokens do not correspond to these same units, is there an issue with using NLM surprisal in more mechanistic accounts of processing? Can surprisal over morphemes reliably predict behavioral results? We replicated previous findings demonstrating a linear relationship between a word's surprisal and its reading time (RT) [10, 13], comparing 5-gram models under orthographic, subword, and morphological segmentation.

Our models were trained on a publicly available portion of the Corpus of Contemporary American English (COCA) [2]. In addition to orthographic words, we used subword tokens under the implementation of Byte-Pair Encoding (BPE) [9] from GPT-2, which fits RTs better than larger models [8]. For morphological units, we used the output of a state-of-the-art segmenter [12]. We then computed word-level surprisal estimates for English RT corpora of eyetracking (Dundee, [6]) and self-paced reading (Natural Stories, [3]) data. Following previous work [10, 13] we fit regression models predicting RTs from surprisal, controlling for word length and frequency, and computed *predictive power* for orthographic, BPE, and morphological surprisal. This measure is the per-token difference in log likelihoods of a surprisal-based model and a model fit to the control predictors, and quantifies how much surprisal improves RT prediction. Our figures report predictive power over held-out test sets under 10-fold cross-validation, following [13]. We found no statistically significant differences between the predictive power of orthographic surprisal and morphological and BPE-based surprisal, suggesting that at least in the aggregate, NLM-based measures in psycholinguistic studies may not be an issue (Fig. 1). We also successfully replicate a major finding in surprisal theory with morphemes, where the individual units are indeed meaningful.

However, this is largely because very few words in the RT corpora were split by the tokenizer in the first place, showing that most of the input to the NLMs (~95% in these corpora) is treated as whole words (Table 1). Looking separately at the set of words split by the BPE tokenizer, we find that the models' predictive power is worse compared to unsplit words. It is likely these words are overall more difficult to predict, since the predictive power of morphological surprisal is also lower (Fig. 2), but that difference is smaller than BPE and not statistically significant. This difference is much clearer with GPT2 surprisal (Fig. 3). Repeating the analysis for words split into multiple morphological units, we do not find major differences between the predictive power of morphological and BPE surprisal (Fig. 4) on monomorphemic and multimorphemic words. In both analyses, we exclude closed-class words to compare the same lexical categories.

Overall, the heuristic offered by NLM-style BPE tokenization might be good enough to capture human processing behavior in the aggregate, but, at the very least, researchers using these models, even for correlational work, should pay attention to tokenization. Recent work [1] claims BPE *is* cognitively plausible since its splits mirror human reaction times in lexical decision tasks, but this largely has to do with form alone, while naturalistic reading involves inferences about meaning. Thus, NLM surprisals can only take us so far in modeling underlying processes of comprehension.

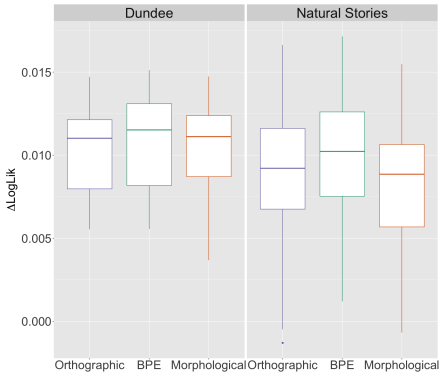

Figure 1: Distribution of predictive power of surprisal under models trained under each segmentation method. The predictive power is positive, replicating previous results [11], and there is no major difference in predictive power associated with each segmentation method relative to orthographic words *(p > 0.05*).

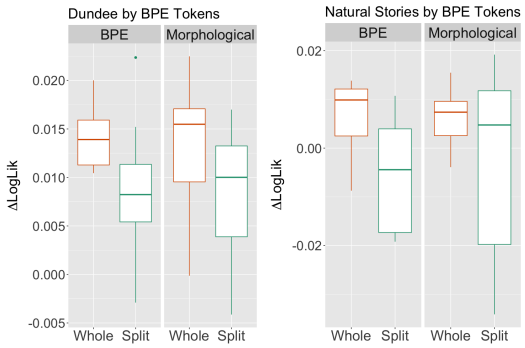

Figure 2: Looking at solely words split into multiple subword tokens, the predictive power of surprisal significantly decreases for the model using BPE tokenization *(p < 0.05)* but not for morphological segmentation *(p > 0.05)*.

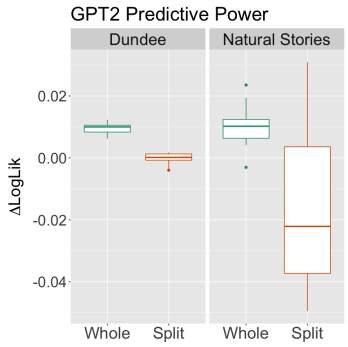

Figure 3: There are statistically significant differences in the predictive power of pretrained GPT2 surprisal between whole and split words using the same BPE tokenizer as the *n*-gram models ($p < 0.001$ for Dundee and $p < 0.01$ for Natural Stories).

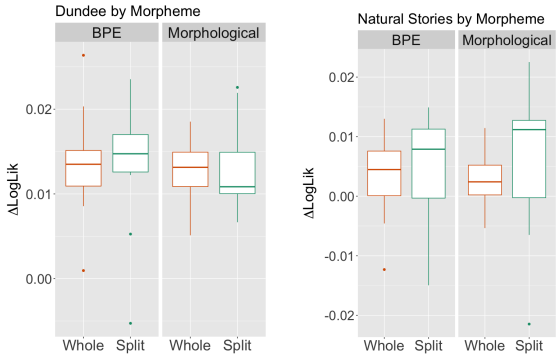

Figure 4: We do not see a comparable effect ($p > 0.05$ for all comparisons) for words split into multiple morphological units.

Table 1: Percent of words in the reading time corpora split by different segmentation methods.

Corpus	Percent Split by BPE Tokenizer	Percent Split by Morphological Segmenter	Percent of Open-Class Words Split by BPE Tokenizer	Percent of Open-Class Words Split by Morphological Segmenter
Dundee	5.6	24.3	11.5	45
Natural Stories	4.8	23.1	10.1	41.7

References: [1] Beinborn & Pinter (*EMNLP* 2023) [2] Davies (*Literary & Linguistic Computing*; 2010) [3] Futrell et al (*LREC* 2018) [4] Gutierrez-Vasques, X., Bentz, C., & Samardžić, T. (*EMNLP* 2023). [5] Gwilliams (*Philosophical Transactions of the Royal Society B* 2020) [6] Kennedy et al (*ECEM* 2003) [7] Oh & Schuler (*TAACL* 2023) [8] Oseki & Marantz (*SciL* 2020) [9] Sennrich et al (*ACL* 2016) [10] Smith & Levy (*Cognition* 2013) [11] Ryskin & Nieuwland (*TiCS* 2023) [12] Wehrli et al (*SIGMORPHON* 2022) [13] Wilcox et al (*Cog Sci* 2020)