Words, Subwords, and Morphemes: What Really Matters in the Surprisal-Reading Time Relationship?

Sathvik Nair and **Philip Resnik**
Linguistics & Institute for Advanced Computer Studies
University of Maryland
{sathvik,resnik}@umd.edu

An important assumption that comes with using neural network-based language models (NLMs) on psycholinguistic data has gone unverified. They are a useful way to quantify lexical predictability during online processing, via computation of surprisal using the estimated probability of a word given context. However, while earlier work used n-gram probability estimates over words (Smith and Levy, 2013), many recent NLMs assign probabilities not to words, but to subword tokens that may not be linguistically meaningful (Sennrich et al., 2016). If a word is split into multiple subunits, one typically computes the word-level surprisal by summing the surprisals of the individual units (Wilcox et al., 2020; Oh and Schuler, 2023, among others), based on the chain rule of probability. Cognitively, this assumes that effort dedicated to a word is the sum of the effort on its parts (Smith and Levy, 2013).

However, human processing involves *morphological* subunits (Gwilliams, 2020). Potential implications of this shift, from words to statistically-determined subword tokens, are an important empirical question. Does the difference in NLMs' units create issues with estimates of NLM-based surprisal in psycholinguistics? And if actual morphemes were the subunits, would models predict behavioral results reliably? We looked at previous findings demonstrating a linear relationship between a word's surprisal and its reading time (Smith and Levy, 2013; Wilcox et al., 2020) through the lens of word segmentation.

We trained 5-gram models using orthographic words, NLM subword units, and morphological subword units, on a publicly available portion of the Corpus of Contemporary American English (COCA; Davies, 2010). We segmented words into NLM subword tokens using the implementation of Byte-Pair Encoding (BPE) from GPT-2, which fits reading times better than larger models (Oh and Schuler, 2023). To obtain morphological units we used a state-of-the-art morphological segmenter

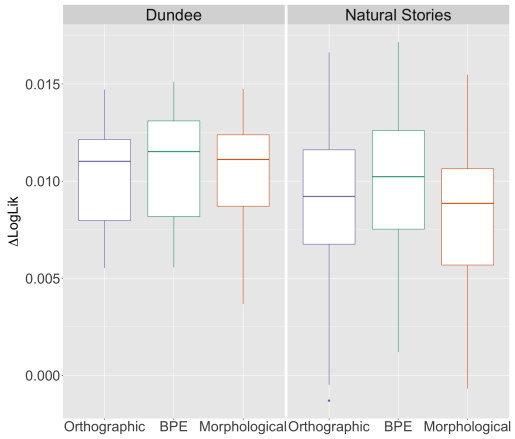

Figure 1: Distribution of predictive power of surprisal under models trained under each segmentation method. There is no major difference in predictive power associated with each segmentation method relative to orthographic words ($p > 0.05$).

(Wehrli et al., 2022). For two English reading time corpora—the Dundee eyetracking corpus (Kennedy et al., 2003) and the Natural Stories self-paced reading corpus (Futrell et al., 2018)—we estimated word-level surprisals, summing over words' subunit surprisals for BPE and morphological segmentation. Following previous work (Smith and Levy, 2013; Wilcox et al., 2020), we fit regression models predicting reading times from surprisal, controlling for word length and frequency. To measure the predictive power of surprisal, we computed $\Delta LogLik$, the per-token difference in log likelihoods of a surprisal-based model and a model fit to the control predictors. Our figures report predictive power over held-out test sets under 10-fold cross-validation, following Wilcox et al. (2020).

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 are likely not to be an issue (Fig 1). However, looking separately at words split and not split by BPE,

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 291-292.
Irvine, California, June 27-29, 2024

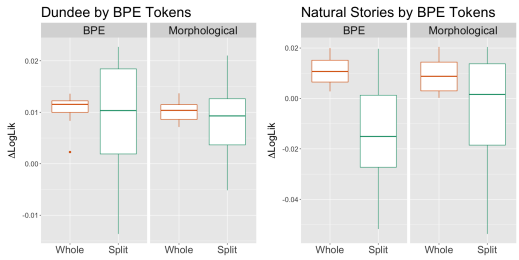

Figure 2: Looking solely at words split into multiple subword tokens, the predictive power of surprisal significantly decreases for the model using BPE tokenization on the Natural Stories corpus relative to unsplit words ($p < 0.001$) and has a much higher variance on the Dundee corpus ($p > 0.05$ for all other comparisons).

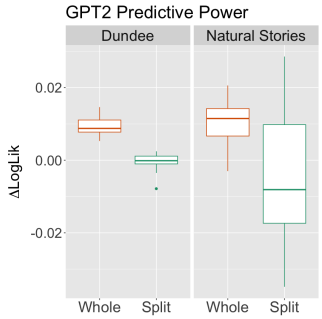

Figure 3: There are statistically significant differences in the predictive power of pretrained GPT2 surprisal between whole and split words ($p < 0.001$ for Dundee and $p < 0.01$ for Natural Stories).

we find predictive power is worse for split versus unsplit words. This suggests good aggregate prediction when using BPE segmentation may largely rely on words BPE *didn't* split up. Split words may be more difficult to predict overall, since the same trend holds for morphological segmentation (Fig 2), but the difference is smaller than BPE and not statistically significant. This difference is much clearer with GPT2 surprisal (Fig. 3).[1]

Our findings suggest it may not be an issue that BPE tokenization often splits up words in a manner that is not necessarily meaningful, at least at the level of measuring word-level reading times in English. However, our outcome with split versus unsplit words suggests a need for further investigation with morphologically rich languages, where predictive power for unsplit words may suffer from sparser counts. In such languages, using morphological subunits may matter more. Our replication here of a key finding in surprisal theory, using morphological subunits for English, is therefore an important, initial verification that estimating surprisal with morphological units can yield psycholinguistically meaningful results.

[1]Repeating the analysis for words split into multiple morphological units and excluding NLTK stopwords (Bird et al., 2009) as a heuristic, we do not see major differences between the predictive power of morphological and BPE surprisal over words with different numbers of morphemes.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Mark Davies. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The natural stories corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Laura Gwilliams. 2020. How the brain composes morphemes into meaning. *Philosophical Transactions of the Royal Society B*, 375(1791):20190311.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

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.

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.

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

Silvan Wehrli, Simon Clematide, and Peter Makarov. 2022. CLUZH at SIGMORPHON 2022 shared tasks on morpheme segmentation and inflection generation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 212–219, Seattle, Washington. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.