Greed is All You Need: An Evaluation of Tokenizer Inference Methods

Omri Uzan[β] **Craig W. Schmidt**[κ] **Chris Tanner**[κμ] **Yuval Pinter**[β]

[β] Department of Computer Science
Ben-Gurion University of the Negev
Beer Sheva, Israel
{omriuz@post, uvp@cs}.bgu.ac.il

[κ] Kensho Technologies
[μ] Massachusetts Institute of Technology
Cambridge, MA, USA
{craig.schmidt, chris.tanner}@kensho.com

Abstract

While subword tokenizers such as BPE and WordPiece are typically used to build vocabularies for NLP models, the method of decoding text into a sequence of tokens from these vocabularies is often left unspecified, or ill-suited to the method in which they were constructed. We provide a controlled analysis of seven tokenizer inference methods across four different algorithms and three vocabulary sizes, performed on a novel intrinsic evaluation suite we curated for English, combining measures rooted in morphology, cognition, and information theory. We show that for the most commonly used tokenizers, greedy inference performs surprisingly well; and that SaGe, a recently-introduced contextually-informed tokenizer, outperforms all others on morphological alignment.

1 Introduction

Modern NLP systems, including large language models (LLMs), typically involve an initial step of mapping raw input text into sequences of subword tokens. These tokens are selected from a large vocabulary of candidates that were produced from algorithms such as Byte-Pair Encoding (BPE; Sennrich et al., 2016), WordPiece (Schuster and Nakajima, 2012), or UnigramLM (Kudo, 2018).

This process, which we refer to as the *inference method* of tokenization, is critical as it determines how all text is represented and subsequently modeled. Each inference method offers distinct mappings, and we assert that it is not well-understood how these methods differ in performance. Furthermore, popular implementation packages such as Huggingface Tokenizers,[1] SentencePiece,[2] and SubwordNMT[3] often obfuscate or even restrict the choice of inference methods, making it unclear if inference-time decoding is compatible with the algorithm used to learn the tokenizer's vocabulary. Moreover, it is yet to be determined whether such a match is ideal, or even necessary.

Tokenizer$_{\text{inference mode}}$	Segmentation
BPE$_{\text{merges}}$	␣Ul tr am od ern
BPE$_{\text{longest prefix}}$	␣Ultra modern
UnigramLM$_{\text{likelihood}}$	␣ U nprecedented
UnigramLM$_{\text{longest prefix}}$	␣Un precedent ed
SaGe$_{\text{longest prefix}}$	␣Inc once iva ble
SaGe$_{\text{likelihood}}$	␣In conceiv able

Table 1: Examples of words being segmented differently by various tokenizers (vocab size 32,000) using different inference modes on the same vocabulary. Each tokenizer's default mode is provided on top.

In Table 1 we present examples demonstrating how the prescribed inference methods of BPE, UnigramLM, and SaGe (Yehezkel and Pinter, 2023) do not necessarily provide the best segmentation for complex English words, even when good segments are available in the vocabulary. BPE's out-of-the-box algorithm merges the cross-morphemic `am` sequence at an early stage, preventing the consideration of `ultra` and `modern` and condemning the downstream model to work with a representation learned for the first-person present form of 'to be'. UnigramLM's ablative algorithm enabled `nprecedented` (which crosses morpheme boundaries) to remain in its final vocabulary of tokens, while SaGe's greedy algorithm masks the boundaries of both the prefix `In` and the suffix `able`. In all cases, an alternative inference method provides a more morphologically-aligned segmentation over the same vocabulary.

Previous work regarding subword tokenization mostly concerns developing vocabulary construction algorithms (Sennrich et al., 2016; Schuster and Nakajima, 2012; Kudo, 2018; Mielke et al., 2021; Yehezkel and Pinter, 2023), finding the optimal

[1] https://huggingface.co/docs/tokenizers
[2] https://pypi.org/project/sentencepiece
[3] https://github.com/rsennrich/subword-nmt

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 813–822

vocabulary size (Gowda and May, 2020; Gutierrez-Vasques et al., 2021), building multilingual vocabularies (Liang et al., 2023), and using space positioning in the vocabulary tokens (Gow-Smith et al., 2022; Jacobs and Pinter, 2022). Others analyze the effects of vocabularies, finding intricate relations between algorithm or vocabulary and downstream performance (Bostrom and Durrett, 2020; Cognetta et al., 2024a), information theory (Zouhar et al., 2023; Cognetta et al., 2024b), cognitive plausibility (Beinborn and Pinter, 2023), impact on society (Ovalle et al., 2024), or morphological alignment (Klein and Tsarfaty, 2020; Hofmann et al., 2021, 2022; Gow-Smith et al., 2024; Batsuren et al., 2024).

Research concerning inference methods has been more scarce, and includes examination of random effects on BPE merges (Provilkov et al., 2020; Saleva and Lignos, 2023) and application of sophisticated search algorithms (He et al., 2020). As far as we know, there exists no comprehensive study comparing inference methods across a variety of vocabularies and sizes using diverse metrics.

In this work, we conduct a controlled experiment isolating the effects of inference methods over four tokenizers, introducing an evaluation suite aggregating intrinsic benchmarks from various theoretical realms.[4] We find that greedy inference methods work surprisingly well for all four vocabularies across morphological and information-theoretic metrics. Furthermore, we demonstrate that SaGe yields state-of-the-art performance according to morphological metrics, and that inference methods that minimize token count perform strongest by cognitive metrics.

2 Inference Methods

Let $\mathcal{V}$ denote a vocabulary of subword tokens and w denote a *word* (or 'pretoken'), the output of a pretokenizer. We define $s(\mathcal{V}, w) := (t_1, ..., t_k)$ as a segmentation of w into k subword tokens such that $\forall i, t_i \in \mathcal{V}$ and that the concatenation of $t_1, ..., t_k$ results in w. We use the term *segmentation* to denote the application of an *inference method* on a text given a *token vocabulary*, as well as its result.

Current widely-employed tokenization schedules couple together the tokenizer vocabulary with the inference method. However, we advocate for decoupling them, as they are independent processes. Specifically, given a fixed token vocabulary produced from pre-training data, one could subsequently use any applicable inference method for the task at hand. Thus, in our experiments, we use various intrinsic metrics to analyze the impact and performance of the several classes of inference methods:

Greedy inference methods only consider and produce one token at each step. We test three greedy approaches: **Longest prefix**, which WordPiece uses by default (Wu et al., 2016), selects the longest token in $\mathcal{V}$ that is a prefix of w, and then continues to iteratively segment the remaining text. **Longest suffix** selects the longest token that is a suffix of w and continues iteratively (Jacobs and Pinter, 2022; Bauwens, 2023). Since this strategy diverges from English Morphology, we consider it an intriguing baseline for assessing the impact of linguistic structure on the inference method. **Longest token** selects the longest token that is contained in w, adds it to the generated segmentation, and then iteratively segments each remaining character sequence. This was proposed by Hofmann et al. (2022) to approximate words by their k longest tokens. They showed that it preserves morphological structure of words and leads to performance gains on some downstream tasks.

Merge rules-based inference methods begin with a word's character sequence and iteratively apply token-forming merge rules learnt by the tokenizer at the vocabulary creation phase, until none can be applied. This is BPE's default inference mode.[5] In our experiments we test two variants for BPE: The **deterministic** merge strategy recursively applies the first applicable BPE merge rule by its order in the trained merge list. **Dropout** (Provilkov et al., 2020) applies each valid merge rule with probability p, leading to a regularization effect where rare tokens surface more often and their embeddings can be better trained. It has been shown to improve machine translation performance.

Likelihood-based inference methods use individual likelihood values assigned to tokens in order to find a segmentation for w where the total likelihood is maximized (Kudo, 2018; He et al., 2020). **Default** uses likelihood values learned during vocabulary construction and considers the likelihood

[4] We release our code and data at https://github.com/MeLeLBGU/tokenizers_intrinsic_benchmark.

[5] While ostensibly also compatible with WordPiece, we found no implementation of the model that provides an ordered list of its merges.

Resource	Type	Size	Reference	License
LADEC	Morphological	7,804	Gagné et al. (2019)	CC BY-NC 4.0 DEED
MorphoLex	Morphological	12,029	Sánchez-Gutiérrez et al. (2018)	CC BY-NC-SA 4.0 DEED
MorphyNet	Morphological	219,410	Batsuren et al. (2021)	CC BY-SA 3.0 DEED
DagoBert	Morphological	279,443	Hofmann et al. (2020)	Not specified—citation based
UniMorph	Morphological	143,454	Batsuren et al. (2022)	CC BY 4.0 DEED
UnBlend	Morphological	312	Pinter et al. (2020)	GPL-3.0
CompoundPiece	Morphological	22,896	Minixhofer et al. (2023)	Not specified—citation based
Cognitive data	Cognitive	55,867	Beinborn and Pinter (2023)	MIT
tokenization-scorer	Information Theory	—	Zouhar et al. (2023)	Not specified—citation based

Table 2: Size, Reference and License details of the resources in our benchmark.

of a segmentation to be the product of individual likelihoods (from which UnigramLM gets its name). **Least tokens** assigns a constant likelihood value to all tokens, effectively selecting a segmentation where the number of tokens is minimized. While not suggested so far as a standalone inference method, this obecjtive is proposed for both vocabulary training and inference in the PathPiece algorithm (Schmidt et al., 2024).

3 Intrinsic Benchmark

Some analyses of tokenizers rely on training language models or translation models and evaluating their performance on downstream tasks. Using this process to isolate effects of tokenization hyperparameters, such as inference method, is both time- and resource-consuming, as well as unstable due to the introduction of multiple sources of randomness throughout the LM/TM pre-training and fine-tuning phases. Few measures have been introduced that are intrinsic to vocabularies and their direct application to corpora, and fewer still avoid conflating the measures with the objectives used in the vocabulary construction process itself. As a result, the body of work focused on improving tokenization schemes is still relatively small.

We create and release a benchmark made to intrinsically evaluate subword tokenizers. We collected word-level datasets and information measures which have been shown, or hypothesized, to correlate with the performance of language models on various downstream tasks. Details on these resources are provided in Table 2. At present, the benchmark is focused on the English language, although corresponding datasets exist for others as well.

Morphological alignment It is commonly assumed that, for a given tokenizer, alignment of word segments to morphological gold-standard segmentations is a predictor of the ability of a language model that uses the given tokenizer to represent words, especially 'complex' ones that are made up of several roots or contain multiple morphological affixes (Schick and Schütze, 2019; Nayak et al., 2020; Hofmann et al., 2021; Gow-Smith et al., 2022). We follow Gow-Smith et al. (2022) and evaluate our tokenizers's alignment with morphological annotations found in LADEC (Gagné et al., 2019), MorphoLex (Sánchez-Gutiérrez et al., 2018), MorphyNet (Batsuren et al., 2021), and DagoBert (Hofmann et al., 2020). We augment these datasets with morpheme segmentation data (Batsuren et al., 2022), novel blend structure detection data (Pinter et al., 2020), and compound separation data (Minixhofer et al., 2023). The number of words in each resource can be found in Table 2. We compare the segmentations generated by the tokenizers with each inference method to gold-standard morphological segmentations using the metric introduced by Creutz and Linden (2004), and report the macro-averaged F_1 score over the different resources.

Cognitive Plausibility We use the benchmark and data from Beinborn and Pinter (2023) to measure the correlation of a tokenizer's output with the response time and accuracy of human participants in a lexical decision task, predicated on the hypothesis that a good tokenizer struggles with character sequences that humans find difficult, and vice versa. We report the average of the absolute value correlation scores across the four linguistic setups (word/nonword $\times$ accuracy/response time).

Tokens distribution statistics We report the Rényi efficiency of different segmentations across a corpus (Zouhar et al., 2023). This measure penalizes token distributions dominated by either very high- and/or very low-frequency tokens, and was shown to correlate strongly with BLEU scores for machine translation systems trained on the respective tokenizers. Recent work (Cognetta et al.,

Vocab	Inference method	Morphological alignment	Cognitive plausibility	Rényi efficiency	Tokens per word	Decoding diff
BPE	longest prefix	.8584	.3266	.4482	1.4273	.0502
	longest suffix	.6467	.3170	.4482	1.4286	.0417
	longest token	**.8738**	.3302	.4474	1.4261	.0484
	least tokens	.7544	.3321	.4476	**1.4237**	.0382
	det. merges	.6309	**.3355**	.4482	1.4308	—
	dropout merge	.6081	.2925	**.4537**	1.5793	.1313
WordPiece	*longest prefix*	**.8488**	.3307	**.4507**	1.4430	—
	longest suffix	.6288	.3198	.4502	1.4435	.0656
	longest token	.8466	**.3332**	.4500	1.4411	.0216
	least tokens	.7342	.3306	.4401	**1.4319**	.0682
UnigramLM	longest prefix	**.9222**	.2858	**.3400**	1.7577	.1187
	longest suffix	.7520	.2690	.2897	1.7624	.0516
	longest token	.8845	.2948	.3040	1.7353	.0406
	least tokens	.8982	**.2953**	.2969	**1.7219**	.0328
	likelihood	.9149	.2937	.2919	1.7314	—
SaGe	*longest prefix*	**.9606**	.2581	**.3217**	1.9445	—
	longest suffix	.7370	.2471	.2832	1.9615	.1704
	longest token	.9236	.2671	.3027	1.9236	.0887
	least tokens	.9125	**.2674**	.2944	**1.8895**	.1318
	likelihood†	.9515	.2664	.2937	1.9156	.1168

Table 3: Intrinsic Benchmark results on a vocab size of 40k. 'Default' decoding algorithms (used in vocabulary construction) in *italics*. Not all methods are applicable to all tokenizers. *Decoding diff* presents the share of pretokens in the MiniPile test set that are differently tokenized using the method, compared with the default. We present correlation scores for performance over the various metric families in Appendix C.
†For SaGe, likelihood is only based on unigram scores obtained before further vocabulary ablation.

2024b) reveals a misalignment between Rényi efficiency and downstream performance in certain cases, reinforcing the necessity of an evaluation suite grounded in diverse domains and disciplines, as advocated in this work. We also measure the average number of tokens per word over a corpus, as a proxy for compression quality (Gallé, 2019). We omit the popular measure of character-length distribution of the tokens in the vocabulary, as it does not vary with segmentation strategy.

Lastly, we report the proportion of pretokens that are segmented different from the default across our reference corpus.

4 Experiments

We evaluate inference methods for the following tokenizer vocabularies: BPE, UnigramLM, WordPiece and SaGe. We use the train split of the MiniPile (Kaddour, 2023) dataset to construct the tokenizer vocabularies. We train vocabularies of sizes 32,768, 40,960, and 49,152, using the HuggingFace Tokenizers library, with identical pre-tokenization, representing the text at byte level. UnigramLM and SaGe require an initial vocabulary for their top-down algorithms; for the former, we used the default implementation of one million top n-grams, while SaGe was initialized with a 262K-size UnigramLM vocabulary. This initial vocabulary also provided us with token likelihood scores for inference, although a more exact implementation would also incorporate the contextual SaGe objective.

Token distribution statistics measurements and decoding diff rates were computed over the test split of the MiniPile dataset. We measure the Rényi efficiency using the tokenization-scorer package[6] with $\alpha = 2.5$. For each tokenizer, all experiments ran within several minutes on a personal laptop computer, highlighting the usefulness of our benchmark as an efficient tool for in-loop hyperparamter tuning.

We present the results on our benchmark for the 40K vocabularies in Table 3. Results for other sizes are presented in Appendix A. A breakdown of individual evaluation subsets is provided in Appendix B.

Inference methods Within each tokenizer, we find that the default ('intended') strategy is often outperformed by others on some measures. We observe a significant difference in morphological alignment when using merge rules-based inference methods. Qualitative analysis showed the findings

[6] https://github.com/zouharvi/tokenization-scorer

illustrated in Table 1, where early merge rules such as 'i-n', 'a-m', or 'o-n' cross morphological boundaries. We notice a similar trend for likelihood-based inference, where frequently-used tokens possess very high likelihood values, sometimes exceeding those of the gold-standard segments. We find that the *least tokens* strategy fares well not only on the token count metric, which is mostly by-design, but also on cognitive measures, suggesting an effect of human preference to minimal word segmentation. Finally, we observe that likelihood-based inference performs poorly in terms of Rényi efficieny, contrary to its stated purpose. *Dropout*, on the other hand, performs well on this measure, in line with its goal. *longest suffix* performs poorly across the board, possibly due to the suffixing nature of the English language, which has complementarily been shown to affect character-level sequential modeling (Pinter et al., 2019). Notably, all our key observations are consistent across vocabulary sizes, as shown in Appendix A.

Inter-tokenizer results Our results align with Bostrom and Durrett (2020)'s finding that BPE is inferior to UnigramLM on morphology alignment. However, we show that some of this gap can be attributed not to the vocabulary but to the inference method. In addition, we find that SaGe is most aligned to morphology by a substantial margin, indicating that its contextualized objective succeeds in retaining meaningful tokens in the vocabulary during ablation. It is important to note that our evaluation is limited to English, a language with relatively low morphological complexity. Previous studies have identified significant tokenization challenges in non-English languages (Mager et al., 2022). Therefore, any definitive conclusions regarding the effectiveness of tokenization methods should ideally encompass a diverse array of languages. BPE and WordPiece, optimized for compression, unsurprisingly perform well above the likelihood-based vocabularies on the information measures. However, we note that this carries over to the cognitive benchmark as well, supporting Beinborn and Pinter (2023)'s findings.

Finally, we note that the two likelihood-based vocabularies follow the exact same within-vocab trends, and those for the two information-based vocabularies are also very close. This highlights the consistency and robustness of our benchmark, although some results are relatively close to each other, which can be expected considering that some inference methods do not change much of the token sequences (see rightmost column of Table 3).

5 Conclusion

In this work, we curated an aggregated benchmark for intrinsic evaluation of subword tokenizers and used it to show the importance of selecting an inference method suited for a vocabulary given a task. Given its computational efficiency, we hope the benchmark can be used in LM training efforts as a fruitful first step to improve tokenization schemes, or to select inference methods on-line. Concretely, our findings suggest that greedy inference is a good choice, especially for morphologically-motivated tasks, even for tokenizers trained on other objectives. Considering its ease of implementation and faster inference, this is an encouraging finding.

In the future, we plan to examine the correlation between our benchmark and various downstream tasks, as well as expand our experimentation to other languages and new algorithms.

Limitations

Our paper contains evaluation of models in the English language. This was done mostly in order to focus this short paper's contribution, and to be able to control for as many possibly-confounding variables such as training data. Nevertheless, a more complete followup would have to include attempts to replicate our findings on other languages, aiming for a set as diverse as possible mostly in terms of typology and script.

Our evaluation is limited to intrinsic measures. While this makes development of tokenizers easier, we acknowledge that the body of work correlating success on these measures with performance of downstream models on end-tasks is incomplete.

Ethical Considerations

Details for human annotation for the cognitive benchmark are documented in the source benchmark's paper (Beinborn and Pinter, 2023), from which we took the data as-is.

Acknowledgments

We would like to thank Charlie Lovering, Varshini Reddy, and Haoran Zhang for comments on early drafts of this paper. We thank the anonymous reviewers for their comments on our submission. This research was supported in part by the Israel Science Foundation (grant No. 1166/23) and by

a Google gift intended for work on *Meaningful Subword Text Tokenization*.

References

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

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.

Thomas Bauwens. 2023. BPE-knockout: Systematic review of BPE tokenisers and their flaws with application in Dutch morphology. Master's thesis, KU Leuven.

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.

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, Online. Association for Computational Linguistics.

Marco Cognetta, Tatsuya Hiraoka, Naoaki Okazaki, Rico Sennrich, and Yuval Pinter. 2024a. An analysis of bpe vocabulary trimming in neural machine translation.

Marco Cognetta, Vilém Zouhar, Sangwhan Moon, and Naoaki Okazaki. 2024b. Two counterexamples to tokenization and the noiseless channel. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16897–16906, Torino, Italia. ELRA and ICCL.

Mathias Creutz and Bo Krister Johan Linden. 2004. Morpheme segmentation gold standards for finnish and english.

Christina L. Gagné, Thomas L. Spalding, and Daniel Schmidtke. 2019. Ladec: The large database of english compounds. *Behavior Research Methods*, 51:2152 – 2179.

Matthias Gallé. 2019. Investigating the effectiveness of BPE: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Edward Gow-Smith, Dylan Phelps, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2024. Word boundary information isn't useful for encoder language models.

Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2022. Improving tokenisation by alternative treatment of spaces. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11430–11443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters

to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2020. Dynamic programming encoding for subword segmentation in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3042–3051, Online. Association for Computational Linguistics.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2020. DagoBERT: Generating derivational morphology with a pretrained language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3848–3861, Online. Association for Computational Linguistics.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. 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 3594–3608, Online. Association for Computational Linguistics.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Cassandra L Jacobs and Yuval Pinter. 2022. Lost in space marking. *arXiv preprint arXiv:2208.01561*.

Jean Kaddour. 2023. The minipile challenge for data-efficient language models. *arXiv preprint arXiv:2304.08442*.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al. 2021. Between words and characters: a brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. CompoundPiece: Evaluating and improving decompounding performance of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 343–359, Singapore. Association for Computational Linguistics.

Anmol Nayak, Hariprasad Timmapathini, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. 2020. Domain adaptation challenges of BERT in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 1–5, Online. Association for Computational Linguistics.

Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. 2024. Tokenization matters: Navigating data-scarce tokenization for gender inclusive language technologies.

Yuval Pinter, Cassandra L. Jacobs, and Jacob Eisenstein. 2020. Will it unblend? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1525–1535, Online. Association for Computational Linguistics.

Yuval Pinter, Marc Marone, and Jacob Eisenstein. 2019. Character eyes: Seeing language through character-level taggers. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 95–102, Florence, Italy. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Jonne Saleva and Constantine Lignos. 2023. What changes when you randomly choose BPE merge operations? not much. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 59–66, Dubrovnik, Croatia. Association for Computational Linguistics.

Claudia H. Sánchez-Gutiérrez, Hugo Mailhot, S. Hélène Deacon, and Maximiliano A. Wilson. 2018. Morpholex: A derivational morphological database for 70,000 english words. *Behavior Research Methods*, 50:1568–1580.

Timo Schick and Hinrich Schütze. 2019. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

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.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Shaked Yehezkel and Yuval Pinter. 2023. Incorporating context into subword vocabularies. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 623–635, Dubrovnik, Croatia. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A Results on Different Vocabulary Sizes

Table 4 presents benchmark results on 32K-sized and 49K-sized vocabularies.

B Detailed Results

Table 5 breaks down the results (for 40K) on individual morphological datasets composing our benchmark. Table 6 Provides the same for individual cognitive measures.

C Inter-Metric Correlations

Table 7 presents the Pearson correlation coefficients between the various intrinsic metrics used in the benchmark. These correlations are calculated based on the aggregated results across all vocabulary sizes.

Vocab	Inference method	Morphological alignment	Cognitive plausibility	Rényi efficiency	Tokens per word	Decoding diff
BPE-32K	longest prefix	.8727	.3122	.4600	1.4511	.0581
	longest suffix	.6496	.3018	.4602	1.4530	.0469
	longest token	.8883	.3152	.4592	1.4498	.0558
	least tokens	.7607	.3174	.4595	1.4469	.0426
	det. merges	.6409	.3201	.4603	1.4551	—
	dropout merge	.6149	.2795	.4656	1.6041	.1316
WordPiece-32K	*longest prefix*	.7819	.3185	.4630	1.4689	—
	longest suffix	.5084	.3089	.4626	1.4698	.0744
	longest token	.7764	.3212	.4622	1.4667	.0243
	least tokens	.7394	.3185	.4508	1.4565	.0769
UnigramLM-32K	longest prefix	.9278	.2855	.3574	1.7803	.1171
	longest suffix	.7610	.2679	.2961	1.7838	.0516
	longest token	.8926	.2930	.3103	1.7534	.0395
	least tokens	.9077	.2937	.3028	1.7418	.0303
	likelihood	.9206	.2931	.2985	1.7501	—
SaGe-32K	*longest prefix*	.9613	.2610	.3454	1.9502	—
	longest suffix	.7449	.2473	.2914	1.9736	.1653
	longest token	.9348	.2685	.3113	1.9319	.0822
	least tokens	.9212	.2691	.3035	1.9084	.1247
	likelihood	.9579	.2679	.3026	1.9246	.1098
BPE-49K	longest prefix	.8440	.3371	.4391	1.4104	.0444
	longest suffix	.6438	.3279	.4390	1.4112	.0379
	longest token	.8637	.3404	.4384	1.4094	.0430
	least tokens	.7464	.3421	.4385	1.4072	.0351
	det. merges	.6208	.3461	.4390	1.4137	—
	dropout merge	.5967	.2996	.4446	1.5610	.1310
WordPiece-49K	*longest prefix*	.7600	.3398	.4413	1.4245	—
	longest suffix	.5133	.3309	.4407	1.4247	.0589
	longest token	.7598	.3421	.4406	1.4228	.0194
	least tokens	.7261	.3401	.4319	1.4145	.0615
UnigramLM-49K	longest prefix	.9157	.2818	.3467	1.7432	.1190
	longest suffix	.7449	.2669	.2849	1.7486	.0516
	longest token	.8750	.2915	.2994	1.7245	.0416
	least tokens	.8908	.2926	.2924	1.7098	.0345
	likelihood	.9095	.2911	.2871	1.7201	—
SaGe-49K	*longest prefix*	.9606	.2566	.3361	1.9414	—
	longest suffix	.7355	.2466	.2783	1.9562	.1735
	longest token	.9200	.2662	.2975	1.9192	.0912
	least tokens	.9053	.2662	.2893	1.8947	.1353
	likelihood	.9455	.2651	.2887	1.9111	.1194

Table 4: Aggregated results on 32K and 49K vocabularies.

Vocab	Inference	Ladec	Morpho-Lex	Morphy-Net	Dago-Bert	Uni-Morph	UnBlend	Compound-Piece
BPE	longest prefix	.9210	.8091	.8511	.8013	.9956	.7404	.8904
	longest suffix	.9497	.6222	.6524	.7116	.0316	.6095	.9502
	longest token	.9147	.8125	.8953	.8618	.9705	.7711	.8905
	least tokens	.9775	.7401	.8303	.8539	.2573	.6489	.9731
	det. merges	.8160	.6781	.6132	.6195	.3233	.6097	.7568
	dropout merge	.7666	.6557	.5871	.5953	.3128	.6213	.7178
WordPiece	*longest prefix*	.9333	.7625	.9114	.8659	.9963	.5569	.9153
	longest suffix	.9447	.6005	.6289	.6844	.1059	.4838	.9535
	longest token	.9275	.7568	.9124	.8765	.9666	.5749	.9112
	least tokens	.9706	.7132	.8253	.8032	.2670	.5897	.9704
UnigramLM	longest prefix	.9551	.8800	.9291	.9087	.9973	.8553	.9299
	longest suffix	.9248	.6387	.8206	.8407	.2777	.8076	.9536
	longest token	.8855	.7534	.9313	.9378	.9135	.8571	.9130
	least tokens	.9660	.8015	.9511	.9593	.7218	.9073	.9801
	likelihood	.9341	.7903	.9645	.9782	.8423	.9205	.9743
SaGe	*longest prefix*	.9734	.9422	.9673	.9600	.9973	.9213	.9626
	longest suffix	.9519	.5996	.7819	.8091	.2403	.8216	.9549
	longest token	.9420	.8390	.9365	.9418	.9711	.8889	.9457
	least tokens	.9856	.8394	.9533	.9632	.7269	.9318	.9877
	likelihood	.9709	.8813	.9809	.9879	.9014	.9492	.9890

Table 5: Results on individual morphological resources.

Vocab	Inference	Words-RT	Words-ACC	nonwords-RT	nonwords-ACC
BPE	longest prefix	−.3136	.4035	.4111	−.1784
	longest suffix	−.3102	.3890	.3987	−.1699
	longest token	−.3164	.4086	.4130	−.1828
	least tokens	−.3146	.4083	.4226	−.1828
	det. merges	−.3285	.4138	.4163	−.1835
	dropout merge	−.2562	.3505	.3908	−.1726
WordPiece	*longest prefix*	−.3198	.4029	.4119	−.1882
	longest suffix	−.3132	.3863	.4028	−.1770
	longest token	−.3226	.4067	.4134	−.1902
	least tokens	−.3146	.4036	.4201	−.1842
UnigramLM	longest prefix	−.2292	.3391	.3920	−.1827
	longest suffix	−.2308	.3235	.3645	−.1572
	longest token	−.2493	.3590	.3904	−.1804
	least tokens	−.2394	.3582	.3978	−.1860
	likelihood	−.2424	.3577	.3926	−.1822
SaGe	*longest prefix*	−.1924	.2896	.3752	−.1754
	longest suffix	−.1895	.2801	.3602	−.1585
	longest token	−.2079	.3047	.3790	−.1767
	least tokens	−.1978	.3034	.3864	−.1821
	likelihood	−.2035	.3043	.3797	−.1780

Table 6: A breakdown of cognitive correlation results across vocabularies and inference methods.

	Morphological alignment	Cognitive plausibility	Rényi efficiency	Tokens per word
Morphological alignment	1	−.5009	−.4799	.5726
Cognitive plausibility	—	1	.6470	−.9588
Rényi efficiency	—	—	1	−.6400
Tokens per word	—	—	—	1

Table 7: Correlations between the different intrinsic metrics.