# MorphBPE: A Morpho-Aware Tokenizer Bridging Linguistic Complexity for Efficient LLM Training Across Morphologies

**Anonymous ACL submission**

## Abstract

Tokenization is fundamental to Natural Language Processing (NLP), directly impacting model efficiency and linguistic fidelity. While Byte Pair Encoding (BPE) is widely used in Large Language Models (LLMs), it often disregards morpheme boundaries, leading to suboptimal segmentation—particularly in morphologically rich languages. We introduce MorphBPE, a morphology-aware extension of BPE that integrates linguistic structure into subword tokenization while preserving statistical efficiency. Additionally, we propose two morphology based evaluation metrics: (i) Morphological Consistency F1-Score, which quantifies the consistency between morpheme sharing and token sharing, contributing to LLM training convergence, and (ii) Morphological Edit Distance, which measures alignment between morphemes and token concerning interpretability. Experiments on English, Russian, Hungarian, and Arabic across 300M and 1B parameter LLMs demonstrate that MorphBPE consistently reduces cross-entropy loss, accelerates convergence, and improves morphological alignment scores. Fully compatible with existing LLM pipelines, $MorphBPE$ requires minimal modifications for integration. The $MorphBPE$ codebase, datasets, and the tokenizer playground will be available with the publication of the work.

## 1 Introduction

Tokenization is a fundamental preprocessing step in NLP, converting raw text into structured units such as bytes (Gillick et al., 2016), characters (Al-Rfou et al., 2019), subwords (Sennrich et al., 2016), words, or multi-word expressions (Gee et al., 2023). Its effectiveness directly influences downstream tasks, as tokenization errors can propagate through the pipeline, impacting overall model performance (Sajjad et al., 2017; Adel et al., 2018). Over the years, tokenization has advanced from basic whitespace-based segmentation to sophisticated statistical and neural approaches (Smit et al., 2014; Otani et al., 2020). In Large Language Models (LLMs), tokenization significantly affects efficiency, context length, and representational accuracy (Dagan et al., 2024). Although tokenization-free architectures have been investigated as potential alternatives (Clark et al., 2022; Deiseroth et al., 2024), most state-of-the-art models—including Gemma (Team et al., 2024), LLaMA (Touvron et al., 2023), DeepSeek (Bi et al., 2024) and OpenAI's GPT series—still rely on Byte Pair Encoding (BPE)-based tokenization for most languages, retaining both its benefits and inherent limitations.

The additive nature of Byte Pair Encoding (BPE) makes it well-suited for concatenative morphology, as seen in English, where morphemes are linearly appended. However, it struggles with non-concatenative morphological systems, such as root-and-pattern morphology in Arabic and Hebrew, where meaning is encoded through non-linear infixation (Khaliq and Carroll, 2013). Similarly, agglutinative languages like Turkish, Hungarian, and Korean pose challenges, as their highly productive affixation processes complicate adherence to morpheme boundaries (Hakkani-Tür et al., 2000). These languages require finer-grained tokenization to preserve linguistically meaningful subword structures. Standard BPE and byte-level tokenization methods often struggle to represent these complex morphological patterns effectively, emphasizing the necessity for morphology-sensitive tokenization approaches that better align with the diverse structural properties of different word formation processes (Marco and Fraser, 2024).

Analyzing BPE output across morphologically rich languages, we observe that its segmentation often disregards meaningful morpheme boundaries, introducing ambiguity and disrupting semantic coherence. For instance, in Arabic, the word الرحمن

(Al-Rahman, "The Merciful") may be incorrectly segmented into من (min, "whom") ال (al, "the") + رح (rah, an incomplete fragment). Here, من (min), a frequent token, is semantically unrelated to the original word, increasing the model's burden in reconstructing meaningful representations. Similar challenges arise in agglutinative and polysynthetic languages, where BPE's greedy merging strategy often fails to align with true morpheme boundaries.

While purely morphology-based segmentation could mitigate these issues, it has also shown limitations in aligning with naturally occurring linguistic patterns in corpus-based learning (Durrani et al., 2019; Marco and Fraser, 2024). Thus, developing tokenization methods that balance morphological integrity with statistical efficiency remains a critical challenge for multilingual NLP.

**Contributions:** We introduce $MorphBPE$, an extension of Byte Pair Encoding (BPE) that integrates linguistic knowledge into subword tokenization. Our key contributions are:

**(i) A Morphology-Aware LLM Tokenizer:** $MorphBPE$ improves adherence to linguistic structures while identifying frequent patterns, balancing *token efficiency* and *interpretability*, particularly in *morphologically rich languages*. It extends BPE by incorporating *morphological structure* while remaining fully compatible with existing LLM training pipelines.

**(ii) Linguistically Informed Tokenizer Evaluation Metrics:** We introduce morphology-aware evaluation metrics to assess tokenization quality:

- **Morph.-Edit Distance Score**: Measures *edit distance* at the *morpheme level*, quantifying segmentation accuracy.
- **Morph.-Consistency F1-Score**: Inspired by (Marco and Fraser, 2024), evaluates the *segmentation consistency*, offering a linguistically grounded metric evaluating whether words that share the same morphemes are also assigned the same tokens, and vice versa.

For benchmarking, we curate a dataset covering diverse morphological typologies (Ge and Comrie, 2022):

- **English**: Fusional, low complexity
- **Russian**: Fusional, moderate complexity
- **Hungarian**: Agglutinative, high complexity
- **Arabic**: Templatic, high complexity

$MorphBPE$ achieves *superior morphological alignment and consistency*, enhancing model interpretability.

**(iii) Empirical Evaluation on LLM Training:** We compare $MorphBPE$ to vanilla BPE on *300M* and *1B* parameter LLMs across the four languages, demonstrating:

- **Lower training loss**, indicating improved linguistic representations.
- **Faster convergence**, enhancing computational efficiency.

By integrating linguistic principles with modern tokenization strategies, MorphBPE bridges the gap between traditional morphological analysis and NLP, providing a computationally efficient and morphologically interpretable tokenization approach for language modeling, particularly in morphologically rich languages like Arabic.

## 2 Related Work

BPE, originally introduced as a text compression algorithm (Shibata et al., 1999), was first adapted for machine translation as a tokenization method in 2016 (Sennrich et al., 2016). Since then, it has become the de facto standard in NLP and Large Language Models (LLMs) due to its efficiency in managing vocabulary size, handling out-of-vocabulary words, and capturing frequent patterns, while offering partial improvements over morphology-based tokenizers (Sennrich et al., 2016).

Despite its widespread adoption, vanilla BPE has several notable limitations: its greedy merging strategy, inefficiencies in cross-lingual settings where similar words with different character variations are not aligned, and inconsistent handling of character-level information across languages. To address these challenges, various extensions have been proposed, including BPE dropout (Provilkov et al., 2020), which introduces stochasticity to improve generalization, sampling-based BPE (Asgari et al., 2019, 2020), which enhances subword diversity, byte-level adaptations (Wang et al., 2020), which aim to improve robustness across scripts, and multilingual BPE variants (Liang et al., 2023), designed to optimize token sharing across languages.

The importance of morphology-aware tokenization for language models has been recognized in several recent studies (Park et al., 2021; Jabbar, 2023; Marco and Fraser, 2024; Weller-Di Marco and Fraser, 2024). However, an integrated solution

that effectively balances morphological information with frequent pattern extraction while remaining fully compatible with modern LLM training pipelines has remained an open problem.

## 3 Methods

Figure 2 provides an overview of our approach. To systematically evaluate **MorphBPE**, we select four languages with distinct morphological typologies, where morphological segmentation is available for training and evaluation at the word level. We determine the vocabulary sizes for each language based on optimal alignment with morphological boundaries. Then we evaluate the vanilla BPE and $MorphBPE$ on the selected vocabulary size using intrinsic metrics detailed in §3.3.

### 3.1 Datasets

#### 3.1.1 Morphological Data

Our dataset comprises morphologically segmented words from four morphologically diverse languages (Ge and Comrie, 2022): English, Russian, Hungarian, and Arabic. The segmentation data for English, Russian, and Hungarian is sourced from the *SIGMORPHON 2022 Shared Task on Morpheme Segmentation* (Batsuren et al., 2022), which provides high-quality morpheme segmentations. To incorporate a *root-based (templatic)* morphological system, we include Arabic, where, we utilize multiple sources: the *Arabic Treebank (ATB)* dataset (Taji et al., 2017), the *Dialectal Segmentation Dataset* (Darwish et al., 2018), and *Quranic morphology data* (Dukes and Habash, 2010). Additionally, we enrich this set with $1M$ high-confidence segmentations of frequent Arabic surfaceforms obtained using *Farasa* (Darwish and Mubarak, 2016). All datasets were cleaned and standardized. Manually annotated segmentations were split into $80\%$ training, $10\%$ validation, and $10\%$ test sets. Table 1 summarizes the dataset composition.

#### 3.1.2 LLM Training Data

For our study on Evaluating $MorphBPE$ vs. BPE Across Languages with Diverse Morphologies: Hungarian, Arabic, Russian, and English, we require a large-scale multilingual training dataset. We selected **FineWeb2** (Penedo et al., 2024), a comprehensive corpus covering over 1,000 languages, to ensure sufficient tokens for training, following the *Chinchilla scaling law* (Hoffmann et al., 2022). This choice enables a balanced token distribution

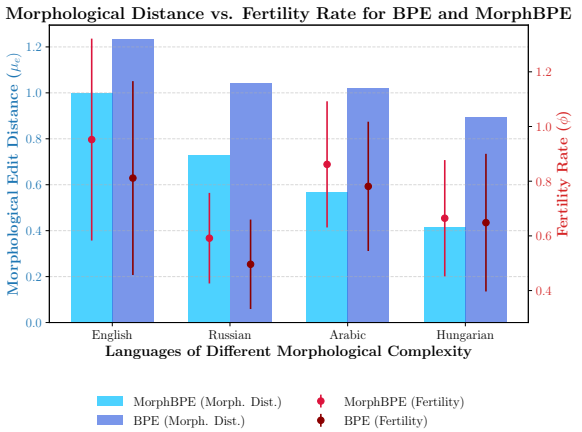


Figure 1: Comparison of morphological distance and fertility rate for BPE and $MorphBPE$ across four languages.

across the selected languages, ensuring fair and robust evaluation of $MorphBPE$ and BPE.

### 3.2 MorphBPE approach

*MorphBPE* is a simple yet effective extension of BPE that prevents frequent symbol pair merges from crossing morpheme boundaries while keeping the rest of the algorithm unchanged (Algorithm 1). This ensures compatibility with standard BPE inference, making $MorphBPE$ easy to integrate into existing pipelines without modifications.

**Algorithm 1** Morphologically-aware Byte Pair Encoding (MorphBPE)

```
Initialize vocabulary with individual characters
Segment the training corpus using morphological segmentation
while number of merges < desired vocabulary size do
    Compute byte-pair frequencies
    Morph-aware Step: Merge the most frequent byte pair without crossing morpheme boundaries
    Update vocabulary with the merged symbol
end while
```

### 3.3 Tokenization Evaluation

Tokenization evaluation can be conducted using intrinsic or extrinsic metrics. Extrinsic evaluation assesses tokenizers in the broader context of LLM performance across diverse capabilities, requiring extensive pre/post training and high-level analysis, which is beyond the scope of this work (Cecchini et al., 2024; Chia et al., 2024). Before evaluating tokenizers in downstream tasks, it is essential to first examine fundamental properties to ensure efficiency and consistency. Therefore, we focus on intrinsic evaluation metrics that provide insights

Table 1: Token Statistics for Morphological Segmentation Datasets Used in BPE and $MorphBPE$ Training and Tokenizer Evaluation Across Languages.

| Language | Morphology Type | # of Words | Avg. Morphemes per Word |
|---|---|---|---|
| Hungarian | Agglutinative | 930,312 | 3.22 |
| Russian | Fusional (moderate complexity) | 784,212 | 3.84 |
| English | Fusional (low complexity) | 571,495 | 2.33 |
| Arabic | Root-based (Templatic) | 1,395,835 | 2.50 |

Table 2: Morph.-consistency evaluation: Precision, Recall, and F1-score for BPE and $MorphBPE$ in different languages. A higher F1-score ($\mu_c$) indicates greater consistency in segmenting words with similar or dissimilar morphemes. Results are reported as mean ± standard deviation over multiple resamples over test sets.

| **Model** | **Precision (Mean ± Std)** | **Recall (Mean ± Std)** | **Morph.-Consistency F1-score ($\mu_c$)** |
|---|---|---|---|
| English BPE (96K) | 0.00 ± 0.00 | 0.03 ± 0.02 | 0.00 |
| English $MorphBPE$ (96K) | 0.20 ± 0.42 | 0.30 ± 0.06 | **0.24** |
| Russian BPE (64K) | 0.10 ± 0.32 | 0.06 ± 0.01 | 0.07 |
| Russian $MorphBPE$ (64K) | 0.69 ± 0.48 | 0.33 ± 0.06 | **0.45** |
| Hungarian BPE (24K) | 0.08 ± 0.25 | 0.29 ± 0.04 | 0.13 |
| Hungarian $MorphBPE$ (24K) | 0.98 ± 0.03 | 0.78 ± 0.07 | **0.87** |
| Arabic BPE (96K) | 0.00 ± 0.00 | 0.08 ± 0.03 | 0.00 |
| Arabic $MorphBPE$ (96K) | 0.89 ± 0.31 | 0.53 ± 0.05 | **0.66** |

into the core characteristics of tokenization in large language models (LLMs).

**(i) Fertility ($\phi$):** Fertility quantifies the number of tokens generated by a tokenizer relative to a baseline, typically a whitespace-based tokenizer (Rust et al., 2021). A lower fertility score generally indicates a more efficient representation, enabling longer contexts. However, this assumption is debatable, particularly for agglutinative languages such as Hungarian and Turkish, where capturing morphological structure necessitates more tokens to provide adequate context for each surface form. As shown in Table 1, languages vary in the average number of morphemes per word. For instance, Hungarian and Arabic require more tokenization compared to English to accurately represent their linguistic structures.

**(ii) Morph.-Edit Distance Score ($\mu_e$):** We introduce a new intrinsic evaluation metric, the *morphological edit distance*, which assesses how well tokenization aligns with the underlying morphological segmentation of words. This metric is computed using a pairwise alignment score based on dynamic programming, ensuring that the order of matching tokens with segmented morphemes is preserved. This approach quantitatively evaluates how effectively a tokenizer respects the morphological structure of the language. We refer to this metric as the Morphology Edit Distance Score ($\mu_e$), which evaluates the interpretability of the tokenizer. While it can be normalized by the number of morphemes in each word, we retain its raw form to provide a clearer indication of the average number of edits required. **(iii) Morph.-Consistency Scores (F1: $\mu_c$):** Inspired by the discussion in (Marco and Fraser, 2024), we propose a morphology consistency measure, which is crucial for language model training. It ensures that words sharing the same morphemes also share tokens (recall score) and that words with shared tokens correspondingly share morphemes (precision score). This evaluation is conducted over a dataset of segmented words, where shared morpheme/token relationships can be treated as either binary events or weighted counts. For simplicity, we adopt a binary scheme, checking whether shared morphemes correspond to shared tokens and vice versa. Since both precision and recall are essential for avoiding unnecessary ambiguity and maintaining a consistent representation of related words, we use their harmonic mean, i.e., the F1-score of morphological consistency, denoted as $\mu_c$.

To ensure practical feasibility given large evaluation datasets, we employ $k$-means clustering ($k = 100$) to group words with similar morphemes and measure scores between $C = 50$ word pairs within each cluster. Precision and recall are estimated through a bootstrapping procedure, drawing $N = 10$ resamples from clusters.

### 3.4 Vocabulary Size Selection

Vocabulary size is a critical hyperparameter in LLM training, directly impacting model perfor-

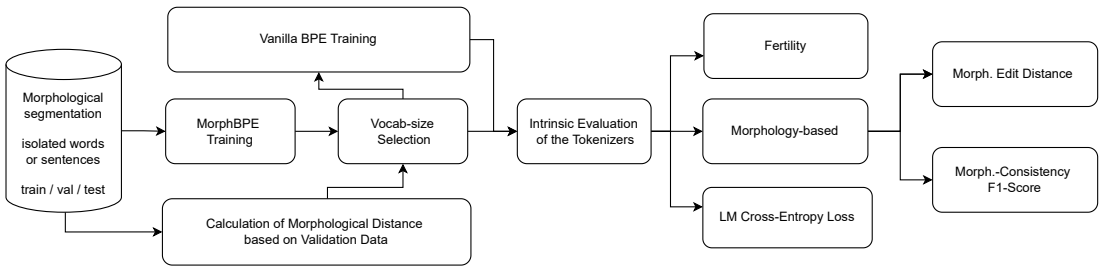


Figure 2: Overview of the $MorphBPE$ study: We evaluate the effectiveness of $MorphBPE$ over vanilla BPE across four morphologically diverse languages (English, Russian, Hungarian, and Arabic) by aligning vocabulary size with morphological segmentation. The we evaluate the tokenizers using the intrinsic evaluation metrics.

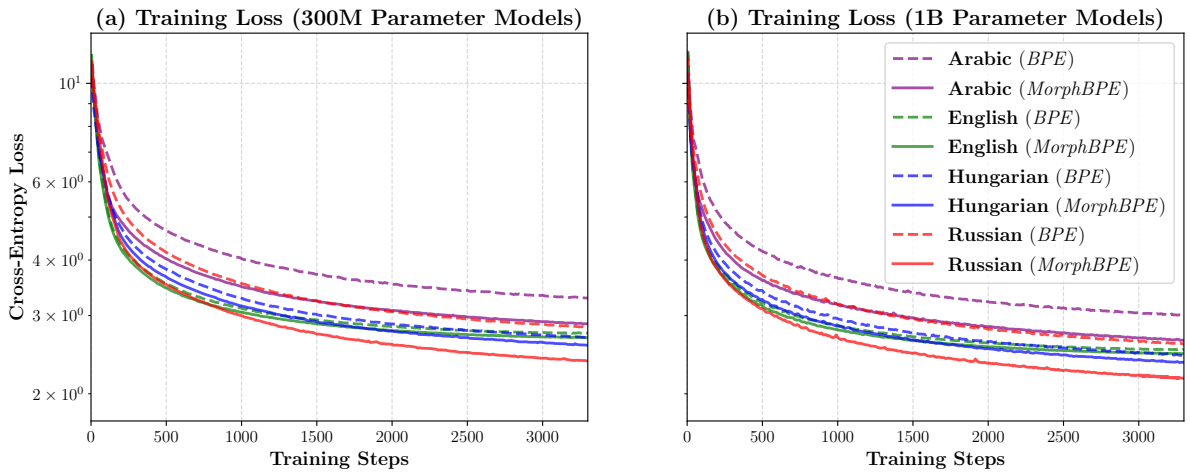


Figure 3: Comparison of training cross-entropy loss between BPE and $MorphBPE$ across four languages. Results are shown for both the small (300M) and large (1B) models.

mance across languages. To determine the optimal vocabulary size in $MorphBPE$, for our four languages, we employed a morphology distance score, $\mu_e$, computed over the development set. We evaluated vocabulary sizes from 8K to 96K in 8K increments, selecting the smallest size beyond which further increases did not yield statistically significant improvements in morphological alignment (*measured via a t-test over the dev. vocabularies*). Through this approach, we determined optimal sizes of **24K for Hungarian** and **64K for Russian**, where larger vocabularies showed diminishing returns. For **English and Arabic**, morphology distance continued improving with larger vocabularies, leading us to select **96K**.

We evaluated the selected tokenizers based on (i) fertility rate ($\phi$), (ii) morphological edit distance score ($\mu_e$), and (iii) morphological consistency score ($\mu_c$) on the test sets of English, Russian, Hungarian, and Arabic. Since fertility rate is a relative measure, we compare both $MorphBPE$ and BPE against a strong multilingual baseline—Bloomz (256K) (Yong et al., 2023), which employs a large vocabulary to accommodate multiple languages. In contrast, $\mu_e$ and $\mu_c$ are directly computed from the test data to evaluate tokenization quality with respect to linguistic structure.

**(iv) Cross Entropy Loss of Language Modeling** ($l_c$) Cross-entropy loss in language modeling measures the divergence between predicted and ground truth outputs. The trajectory of training cross-entropy loss indicates how quickly a model converges and improves next-token prediction. This metric is closely related to model perplexity, a standard intrinsic evaluation measure for language models. However, cross-entropy loss is only comparable across models with identical vocabulary sizes, as vocabulary variations directly affect the model's branching factor.

### 3.5 Language Model Training

To assess the scalability of our approach, we trained two model sizes—**300M (small)** and **1B (large)**—using decoder architectures within the

LLaMA-Factory framework (Zheng et al., 2024). For each language, we trained models with both vanilla BPE and $MorphBPE$ of the same vocabulary sizes, resulting in four models per language. Training loss was monitored and compared across languages and tokenization methods to evaluate their impact on learning efficiency. We ensured passing $\approx 6B$ tokens to the small and $\approx 20B$ tokens to the large model compatible with the *Chinchilla scaling law* (Hoffmann et al., 2022).

## 4 Results

### 4.1 Morphological Metrics and Fertility

The results in Figure 1 and Table 2 show a clear trend: *MorphBPE* consistently achieves lower morphological edit distance ($\mu_e$) and higher morphological consistency ($\mu_c$) compared to BPE, with a slight increase in fertility rate across all languages. The extent of improvement varies based on the morphological complexity of the language. The gap in $\mu_e$ and $\mu_c$ between *MorphBPE* and BPE is larger for Hungarian and Arabic, which have more complex morphological structures. These results indicate that *MorphBPE* better preserves linguistic structure, particularly in morphologically rich languages, while BPE tends to over-fragment words based on subword frequency rather than morpheme boundaries. Higher $\mu_c$ of *MorphBPE* also reflects consistent tokenization which morphology, which can impact the convergence of language model training.

### 4.2 Training cross-entropy loss

The training cross-entropy loss for the four languages, using the same vocabulary and comparing BPE and $MorphBPE$, is presented in Figure 3. The results are shown over a training window of $\approx 14B$ tokens for both small and large models, with the selected interval chosen for clarity, as the overall trend remains consistent throughout training. The results indicate that $MorphBPE$ consistently improves cross-entropy loss across all languages and model sizes, even for English language. This improvement is particularly pronounced in morphologically richer languages, where the reduction in loss is more significant. The results demonstrate lower training loss, indicating improved linguistic representations as well as faster convergence.

## 5 Discussions and Conclusion

In this work, we introduced $MorphBPE$, a morphology-aware extension of BPE that integrates linguistic knowledge into subword tokenization. Through extensive empirical evaluation across English, Russian, Hungarian, and Arabic, we demonstrated that $MorphBPE$ consistently enhances LLM training efficiency by reducing cross-entropy loss, improving morphological alignment, and accelerating convergence across both 300M and 1B parameter models.

Another key contribution of this work is the introduction of linguistically informed tokenizer evaluation metrics, addressing a critical gap in current tokenization evaluation. The Morphological Consistency F1-Score provides a structured measure of segmentation stability, which is essential for ensuring consistent morpheme-level representations during LLM training. This stability directly contributes to better generalization and improved learning efficiency, particularly for morphologically rich languages. Meanwhile, the Morphological Alignment Score, based on edit distance at the morpheme level, serves as a linguistically grounded metric, that can contribute to the interpretability of the tokenizer.

We show that $MorphBPE$, despite having higher fertility, results in a more interpretable and more consistent and more efficient tokenizer for LLM training. This suggests that fertility—a commonly used metric in tokenization evaluation—may not be the most reliable indicator of tokenizer quality of an efficient LLM training.

An additional advantage of $MorphBPE$ is its full compatibility with existing LLM training and inference pipelines, requiring minimal modifications to the tokenization process. This ensures easy integration without disrupting standard workflows. Furthermore, an efficient implementation of $MorphBPE$ training and evaluation metrics will be released with this work, enabling reproducibility and facilitating further research in morphology-aware tokenization.

## 6 Limitations

We demonstrated the effectiveness of $MorphBPE$ across four languages with diverse morphological typologies. However, future work can extend this evaluation to additional languages. One limitation is that $MorphBPE$ relies on the availability of morphological segmentation

data, which is not yet accessible for all languages. Efforts such as UniMorph (Kirov et al., 2018) and MorphyNet (Batsuren et al., 2021) are helping bridge this gap, but further development is needed. Additionally, an important next step is the extrinsic evaluation of LLMs trained with $MorphBPE$, assessing their impact on higher-level capabilities.

## References


Heike Adel, Ehsaneddin Asgari, and Hinrich Schütze. 2018. Overview of character-based models for natural language processing. In *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part I 18*, pages 3–16, Cham. Springer International Publishing.

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3159–3166.

Ehsaneddin Asgari, Alice C McHardy, and Mohammad RK Mofrad. 2019. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). *Scientific reports*, 9(1):3577.

Ehsaneddin Asgari, Masoud Jalili Sabet, Philipp Dufter, Christopher Ringlstetter, and Hinrich Schütze. 2020. Subword sampling for low resource word alignment. *arXiv preprint arXiv:2012.11657*.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

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.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

David Cecchini, Arshaan Nazir, Kalyan Chakravarthy, and Veysel Kocaman. 2024. Holistic evaluation of large language models: Assessing robustness, accuracy, and toxicity for real-world applications. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 109–117, Mexico City, Mexico. Association for Computational Linguistics.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2024. InstructEval: Towards holistic evaluation of instruction-tuned large language models. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 35–64, St. Julian's, Malta. Association for Computational Linguistics.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Gautier Dagan, Gabriele Synnaeve, and Baptiste Rozière. 2024. Getting the most out of your tokenizer for pre-training and domain adaptation. *arXiv preprint arXiv:2402.01035*.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate Arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1070–1074, Portorož, Slovenia. European Language Resources Association (ELRA).

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect Arabic POS tagging: A CRF approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Björn Deiseroth, Manuel Brack, Patrick Schramowski, Kristian Kersting, and Samuel Weinbach. 2024. T-FREE: Subword tokenizer-free generative LLMs via sparse representations for memory-efficient embeddings. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21829–21851, Miami, Florida, USA. Association for Computational Linguistics.

Kais Dukes and Nizar Habash. 2010. Morphological annotation of Quranic Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. 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 1504–1516, Minneapolis, Minnesota. Association for Computational Linguistics.

Pingping Ge and Bernard Comrie. 2022. Correlations of valency alternations and morphological types: A typological perspective. *Lingua*, 273:103304.

Leonidas Gee, Leonardo Rigutini, Marco Ernandes, and Andrea Zugarini. 2023. Multi-word tokenization for sequence compression. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 612–621, Singapore. Association for Computational Linguistics.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1296–1306, San Diego, California. Association for Computational Linguistics.

Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2000. Statistical morphological disambiguation for agglutinative languages. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

J. Hoffmann, S. Borgeaud, M. Arthur, E. Buchatskaya, T. Cai, E. Rutherford, D. Casas, L. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. Rae, O. Vinyals, and L. Sifre. 2022. training compute-optimal large language models. *Arxiv*.

Haris Jabbar. 2023. Morphpiece: Moving away from statistical language representation. *arXiv preprint arXiv:2307.07262*.

Bilal Khaliq and John Carroll. 2013. Induction of root and pattern lexicon for unsupervised morphological analysis of Arabic. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1012–1016, Nagoya, Japan. Asian Federation of Natural Language Processing.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Marion Di Marco and Alexander Fraser. 2024. Subword segmentation in LLMs: Looking at inflection and consistency. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12050–12060, Miami, Florida, USA. Association for Computational Linguistics.

Naoki Otani, Satoru Ozaki, Xingyuan Zhao, Yucen Li, Micaelah St Johns, and Lori Levin. 2020. Pre-tokenization of multi-word expressions in cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4451–4464, Online. Association for Computational Linguistics.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. Fineweb2: A sparkling update with 1000s of languages. *HuggingFace*.

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.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual 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 3118–3135, Online. Association for Computational Linguistics.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, Ahmed Abdelali, Yonatan Belinkov, and Stephan Vogel. 2017. Challenging language-dependent segmentation for Arabic: An application to machine translation and part-of-speech tagging. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 601–607, Vancouver, Canada. Association for Computational Linguistics.

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.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching. *Technical Report DOI-TR-161, Department of Informatics, Kyushu University*.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 21–24, Gothenburg, Sweden. Aalto University, Association for Computational Linguistics.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

Marion Weller-Di Marco and Alexander Fraser. 2024. Analyzing the understanding of morphologically complex words in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1009–1020, Torino, Italia. ELRA and ICCL.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.