Evaluating Portuguese Tokenizers as Morpheme Sequence in Relation to LLM Downstream Performance

Guilherme L. Mello[1,2], **Marcelo Finger**[1,2]

[1] Center for Artificial Intelligence – C4AI/USP

[2]Departamento de Ciência da Computação
Instituto de Matemática, Estatística e Ciência da Computação
Universidade de São Paulo – IME/USP

{guilhermelmello,mfinger}@ime.usp.br

Abstract. *Sub-word tokenizers allow us to handle open vocabulary problems using a relatively small set of tokens. Despite its ease of use, it usually relies on a data-driven approach that does not directly employ linguistic or morphologic features for text tokenization. [Bostrom and Durrett 2020] and [Hofmann et al. 2021] demonstrate that morphemes improve LLM performance on English texts. In this work, we explore the hypothesis that tokenizers that are capable of producing a token sequence better aligned with a morpheme sequence can improve LLMs performance on Brazilian Portuguese. To evaluate how the presence of morphemes can impact LLMs for Brazilian Portuguese, we propose MorphEval-PT, a new evaluation procedure based on the psycholinguistic concept of morphological models of word processing. We build new BPE and Unigram vocabularies that are evaluated on MorphEval-PT and, in order to validate the impact of morphemes on LLMs performance, train new LLMs from scratch and evaluate its performance on downstream tasks. Consistently, BPE demonstrates a higher precision score than Unigram in its ability to represent morphemes, as well as better performance on every downstream task. These promising results indicate that accessing the ability of tokenizers to represent morphemes is an important feature in the development of LLMs for Brazilian Portuguese and that MorphEval-PT is a good and lightweight method to improve LLM performance before any pre-training.*

1. Introduction

Different tokenization methods can be used to pre-process the input of a Large Language Models (LLMs) based on *Transformers* [Vaswani et al. 2017], ranging between word, sub-word and character level. Word level tokenization creates an extensive but closed vocabulary, resulting in several unknown words. Also, applying word level tokenizers on languages like Japanese and Chinese, when words do not have a clear boundary like white spaces, represents an extra challenge. Sub-word and character level tokenizers allow the construction of an open vocabulary, reducing the problem of Out Of Vocabulary (OOV) words by merging sub-word tokens. As a side effect of the small vocabulary, character level tokenizers create a long sequence of tokens. By including words, sub-words and characters, sub-word level tokenizers combine the flexibility of an open vocabulary with a not so long sequence of tokens. Besides its interesting features, the main sub-word level tokenizers are data-driven and create many tokens that do not represent any grammatical

or semantical information. In any language, the meaning of words emerge from their morphemes, the smallest semantical units, and their combination follows rules of word formation [Gonçalves 2019, Plag 2018, Leminen et al. 2019]. This semantical structure will be explored in this paper.

The literature show evidences that morphology have a good impact on LLMs performance [Bostrom and Durrett 2020, Hofmann et al. 2021, Li et al. 2019, Hou et al. 2023, Park et al. 2021]. As opposed to [Bostrom and Durrett 2020], we show that BPE slightly outperform Unigram on every downstream task. We hypothesize that this gain is related with its capability of creating a more morphological coherent sequence of tokens. The main contribution of this paper are: (a) development of MorphEval-PT, a method to evaluate how likely a tokenizer can resemble a sequence of morphemes; (b) a fair comparison between Byte-Pair Encoding and Unigram tokenizers; and (c) present evidences that a morphologically coherent sequence of tokens can improve LLMs performance on downstream tasks.

In Section 2 we provide a short discussion about related works and an overview about the tokenizers algorithms and psycholinguistic models of how the brain is capable of processing the meaning of words. Section 3 describes MorphEval-PT, the evaluation framework proposed in this paper. Sections 4.1 and 4.2 describe the datasets used to train and evaluate LLMs, followed by the experiment setup in Section 5. We conclude this paper with a discussion in Section 6.

2. Related Work

Sub-word level tokenizers have become a consolidated option [Devlin et al. 2018, Radford et al. 2018, He et al. 2021, Yang et al. 2025], like WordPiece [Schuster and Nakajima 2012], Byte-Pair Encoding [Sennrich et al. 2016] e Unigram [Kudo 2018]. When introducing Unigram, [Kudo 2018] could not obtain better result than BPE on a direct comparison. But Unigram enabled the calculation of probabilities and the usage of a regularization method that improved the performance against BPE. [Bostrom and Durrett 2020] shows that BPE is a suboptimal option, assigning Unigram's better performance to its capability to create a more morphological coherent tokenization. Even though sub-word level tokenizers are efficient on generating a compact vocabulary that can handle rare and unknown words, they are still based on statistical patterns, creating several tokens that do not carry any grammatical or semantic information. In a different context, [Li et al. 2019] apply character level tokenizer on Chinese texts, a language that symbols already carry a good amount of semantic information, showing evidences that a morphological coherent sequence of tokens can improve LLMs performance.

[Hofmann et al. 2021] analyzes the influence of tokenizers on LLM's ability of interpreting complex english words. For that, they used psycholinguistic theories that try to explain how the human brain is capable of processing the meaning of complex words in a mental lexicon [Plag 2018, Leminen et al. 2019]. To improve BERT's interpretation skills, they explored derivation, a word formation process, while keeping its original vocabulary and weights. Replacing only the WordPiece algorithm by a tokenizer that generated a sequence of tokens that favored the use of the morphemes already present on the vocabulary, they allowed BERT to consume a sequence of tokens that was bet-

ter aligned with a sequence of morphemes, creating better semantic representations that improved downstream tasks performance. They also show that the same text can be understood in different ways by LLMs when considering a different sequence of tokens. [Hou et al. 2023] and [Park et al. 2021] also show the importance of considering morphology when tokenizing a text.

The main morphological resources available for Portuguese are based on Finite State Morphology [Beesley and Karttunen 2003], like Port4Nooj [Mota et al. 2016] and MorphoBr [de Alencar et al. 2018]. Even though they are linguistically precise, they are more complex and do not provide a direct sequence of morphemes. Also, they focus on inflectional morphology and do not provide enough information about derivation processes. MorphyNet [Batsuren et al. 2021] is an alternative resource that provides an easy to use dataset that contains morphological information and a morpheme segmentation extracted from Wiktionary. The evaluation of morpheme segmentation can be handled as a boundary detection problem, commonly measured by metrics like precision, recall and F1 Score [Creutz and Linden 2004, Nouri and Yangarber 2016]. We present MorphEval-PT (Section 3) by adapting MorphyNet's dataset and including the psycholinguistics concept of morphological processing in our evaluation setup.

2.1. Tokenizers

Tokenization algorithms can range between word, sub-word and character levels. Sub-word tokenizers represent a middle ground between word and character levels that combine their best features: the open and small sized vocabulary that generates a not so long sequence of tokens. In this paper we consider Unigram [Kudo 2018] and BPE [Sennrich et al. 2016] as tokenization methods. WordPiece [Schuster and Nakajima 2012] is a well known tokenization algorithm, but was never openly published by Google. Available implementations are based on public information[1]. Thus, WordPiece is usually not considered by the literature. Accordingly, we leave WordPiece out of our experiments.

BPE was introduced to solve an open vocabulary problem with a fixed sized set of sub-word tokens to significantly decrease the occurrence of OOVs. The BPE vocabulary is built in a bottom-up setup, initialized with basic symbols, and the training corpus is converted in a sequence of single character tokens. During training, the most frequent pair of tokens are iteratively merged to create a new token in the vocabulary, and the merge rule is stored in an ordered list. The training repeats until the vocabulary reaches the desired size. On text tokenization, BPE deterministically applies the merge rules in the same order as they were learned. This creates a compact sequence of tokens that proved valuable for LLMs.

Unigram follows a different strategy than BPE, building its vocabulary in a top-down approach. The vocabulary is initialized with a large set of tokens that is reduced iteratively until it reaches the required size. The vocabulary can be initialized with all possible tokens or an heuristic algorithm can be used, like Enhanced Suffix Array. At each step, the probability $p(x_i)$ is calculated for every token i. The $loss_i$ represents how much loss is introduced when the token i is removed from vocabulary. At each step, a percentage of the least meaningful tokens is removed from the vocabulary. On text

[1]WordPiece implementation: https://huggingface.co/learn/llm-course/chapter6/6

tokenization, the Viterbi algorithm can be used to find the most likely sequence of tokens x^* in the set of probable tokenizations $S(X)$.

2.2. Word Formation and The Mental Lexicon

In linguistics, morphology studies the words forms, covering the analysis of its internal structure and the processes of word formation [Gonçalves 2019, Plag 2018]. Morphemes are the smallest semantic units and, when combined, can generate words with more complex meanings. For example, in the word `morfologia`, the meaning "studies related to forms" emerges when combining the meaning of its morphemes `[morf(o)]` and `[logia]`, that mean "form" and "study", respectively. Even though morphemes are recurring parts in a language, what make them reusable is its semantic relevance [Gonçalves 2019]. Thus, words like *alho* and *bugalho*, despite having a recurring part, they do not share a morpheme as it does not have a semantic relation. The main morphological operations are derivation and inflection. Inflection creates different forms of the same lexical item, changing aspects like gender, number, tense, mood or person without changing its meaning. Derivation creates new lexical items, with new and more complex meanings. In this paper, we explore inflection and derivation by affixation because of its concatenative characteristic that resembles how sub-word tokenizers split text into tokens.

In psycholinguistics, the models of morphological process try to explain how the human brain is capable to process language and to understand the meaning of words [Plag 2018, Leminen et al. 2019]. These models contains two different routes for word processing: storage and decomposition routes. The meaning of words that are frequently used can be more efficiently recovered when stored as a single unit in the mental lexicon by the storage route. In contrast, less frequent words are more efficiently recovered when the meaning is computed by combining the meaning of its morphemes (decomposition route). [Hofmann et al. 2021] shows that it is possible to relate these mechanism with how sub-word tokenizers split text into tokens, even when tokens do not represent morphemes.

3. MorphEval-PT

Evaluating the alignment of a token sequence as a morpheme sequence is not an easy task for Portuguese, a language in which resources are limited. To partially fill this gap, we present MorphEval-PT[2], an evaluation method based on psycholinguistic models of morphological processing to access how likely sub-word tokenizers can generate a morphologically coherent sequence of tokens. MorphEval-PT is divided in two components: (i) dataset and (ii) evaluation metrics. As Portuguese does not provide the required resources to achieve the goal of this paper, we adapted MorphyNet [Batsuren et al. 2021]. Next, we introduce MorphyNet, discuss its limitations and how we handled that to build MorphEval-PT's dataset and evaluation metrics.

3.1. MorphyNet Dataset

MorphyNet [Batsuren et al. 2021] is a multilingual dataset that provides information about inflectional and derivational morphology. For inflectional morphology, it provides the lemma, inflected form, morphological features and a morpheme segmentation, as exemplified in Table 1. The derivational morphology provides the source word, target word, POS tags, a single morpheme and its type (prefix or suffix), as exemplified in Table 2.

[2]Available at: https://github.com/guilhermelmello/MorphEval-PT

Lemma	Inflected Form	Morphological Features	Morpheme Segmentation
infiltrar	infiltrar	V;1;SG;NFIN	-
infiltrar	infiltrares	V;NFIN|2;SG	infiltrar|ares
infiltrar	infiltrados	V|V.PTCP;PST|MASC;PL	infiltrar|ado|s
infiltrar	infiltradas	V|V.PTCP;PST|FEM|PL	infiltrar|ado|a|s

Table 1. Example of Portuguese inflectional information provided by MorphyNet.

Source Word	Target Word	Source POS	Target POS	Morpheme	Type
moral	amoral	J	J	a	prefix
adição	adicionar	N	V	ar	suffix
jurar	juramento	V	N	mento	suffix
haver	reaver	V	V	re	prefix

Table 2. Example of Portuguese derivational information provided by MorphyNet.

3.2. Dataset Construction and Limitations

To guarantee a better data quality, we started by removing entries where words are not present in a Brazilian Portuguese dictionary. For that, we used Enchant/Hunspell[3][4] spell checkers, that use the dictionary provided by VERO Project[5]. In this step, we kept the entries where the lemma (source word) or inflected form (target word) were validated by the spell checker. To be able to evaluate sub-word tokenizers, the MorphEval-PT dataset must contain an input word and its expected morpheme segmentation. Considering the inflectional data provided by MorphyNet, it is possible to use the inflected form as input and the morpheme segmentation as output. However, as Table 1 shows, some limitation exists: (i) only inflectional morphemes are available and (ii) the morphemes are expressed in their canonical form, which cannot recreate the inflected form only by concatenation. In respect to derivational data, the target word, morpheme and its type can be used. Also, some limitations exits: (i) only the morpheme used to create the derived word is provided by MorphyNet (see Table 2).

To obtain a sequence of morphemes that recreates the input word, different rules were applied for inflection and derivation. For inflectional data we considered the inflected form as input and transformed the sequence of morphemes from underlying to surface level. Entries that does not provide a morpheme segmentation were also removed. To create the surface level, we converted the feminine morphemes `ado|a` into `ada`. Next, we ignored the first morpheme as it represents the lemma and, using the remaining morphemes in reverse order, iteratively removed them from the inflected form. As an example, `infiltrar|ado|a|s` was converted into `infiltr|ada|s`, which concatenation results in *infiltradas* as expected. Converting derivational data was simpler and only required the removal of the affix from the derived form based on its type (prefix or suffix), resulting in a segmentation with two morphemes. For example, the words *amoral* and *adicionar* were segmented as `a|moral` and `adicion|ar`, respectively. For derivational morphology we also kept the affix type to assist during evaluation. Both derivational and inflectional data were validated to guarantee that input word could be ob-

[3] https://rrthomas.github.io/enchant/
[4] https://github.com/hunspell/hunspell
[5] https://pt-br.libreoffice.org/projetos/vero

tained by concatenating the morphemes. Any entry that did not respected it was removed from MorphEval-PT dataset. The final dataset contains 11501 derivation and 299807 inflection entries, representing 97.68% and 79.66% of MorphyNet's entries, respectively.

With MorphEval-PT's dataset in place, some usage limitations must be considered. As no data balancing were applied, some morphemes like *anti-*, *-mente* and *-s* are more frequent than others and their higher frequency can benefit tokenizers that are better at their token type. Another limitation concerns the segmentation provided by MorphyNet. As inflectional data provides only inflectional morphemes, existing derivational morphemes are not specified, like *desempates* and *prefabricados* which are segmented as `desempate|s` and `prefabric|ado|s`, respectively. For derivational data, only a single derivational morpheme is provided by MorphyNet and entries that contain more than one morpheme may appear multiple times as different entries. The word *descolonização*, for example, have different entries for each affix, providing `des|colonização` and `descoloniza|ção` as segmentations. Even though some words can be fully segmented by combining multiple entries, to evaluate the extent of this feature requires specialized knowledge and is beyond of the scope of this paper. Thus, a proper usage of this dataset is limited to affixes identifications rather than a complete morpheme segmentation.

3.3. Evaluation Metrics

To evaluate a sequence of morphemes we considered it as a boundary detection problem [Creutz and Linden 2004]. In this setup, we try to evaluate if a sequence of tokens have the same boundaries as the provided morphemes, which can be framed as a binary classification problem. For example, the targets `amoral|ismo` and `infiltr|ada|s` can be translated in the sequence of labels $[0, 0, 0, 0, 0, 1, 0, 0, 0]$ and $[0, 0, 0, 0, 0, 1, 0, 0, 0, 1]$, respectively. To handle dataset limitations, as the morphemes may not represent a complete morphological segmentation and the correctness of the base of the complex word can not be assured, we removed their boundaries from target labels. Taking the previous examples, as `amoral|ismo` and `infiltr|ada|s` does not provide the segmentation of their prefixes, we only considered `|ismo` and `|ada|s`, resulting in the label sequences $[1, 0, 0, 0]$ and $[1, 0, 0, 0, 1]$, respectively. The same is applied when the base of the word is placed on the right side of the sequence, then the segmentation `anti|moralismo` expects the sequence $[0, 0, 0, 1]$ as target labels. When relating this labels with psycholinguistics models of morphological processing, they represent the decomposition route. But sub-word tokenizers can also segment a word as a single token, which represents the storage route, have a clear semantical meaning and is also morphological coherent. To mimic both routes of the psycholinguists models this is also considered as correct segmentation by MorphEval-PT's evaluation metrics (precision, recall and F1 Score).

In the context of morphological coherent segmentation, $precision = \frac{TP}{TP+FP}$ measures how many of the boundaries created by the tokenizer were correctly placed. A higher precision indicates a higher morphological coherence while a lower precision shows that tokens are smaller units than morphemes. Two morphemes can be merged in a single token and still have a clear semantical meaning, as a morphological coherent token; thus, precision is the main metric to evaluate morphological coherent tokenization. The $recall = \frac{TP}{TP+FN}$ measures how many boundaries were correctly recovered by the tokenizer, indicating if at least the morpheme boundaries were correctly placed. As it

does not consider every boundary created by the tokenizer, character level tokenization can result in 100% recall. Therefore, recall can only show us evidences if morphemes are likely to be merged in a single token or not. The $F_1 = 2\frac{Precision \times Recall}{Precision + Recall}$ is the harmonic average between precision and recall.

4. Methods

4.1. Corpus Carolina

In order to pretrain Brazilian Portuguese LLMs from scratch we used Corpus Carolina [Crespo et al. 2023] in its 2.0.1 version, which is an open corpus with curated texts extracted from web. It includes 2108999 texts spanning different domains, like legal texts, wikis and social media. The corpus is available for download on Hugging Face[6].

4.2. Downstream Tasks

In this paper we used ASSIN [Fonseca et al. 2016], ASSIN 2 [Real et al. 2020] and HateBR [Vargas et al. 2022] to evaluate LLMs on downstream tasks. ASSIN and ASSIN 2 contain 10000 and 9448 a pair of sentences, respectively, split in training, validation and test sets that were manually annotated for Recognizing Textual Entailment (RTE) and Semantic Textual Similarity (STS). Its texts were extracted from news and contains both Brazilian and European Portuguese. STS is a regression problem and its labels range from 1 to 5, indicating the similarity level between the sentences. RTE is a classification problem with different label set for each dataset. ASSIN provide the labels *entailment*, *paraphrase* and *none*. ASSIN 2 simplified the RTE task to a binary classification by using *entailment* and *non-entailment* labels.

HateBR [Vargas et al. 2022] is an open corpus motivated by the growing hostility in social media developed to help research about offensive and hate speech detection. It contains 7000 comments extracted from social that were manually annotated in 3 levels of hostility. The fist annotation level represents the offensive language classification task, a binary classification problem that contains 3500 comments with some type of offensive language. The second annotation level if a fine-grained annotation for offensive comments (3500) that specifies a level of offensiveness: highly, moderately and slightly. The third level of annotation represents different types of hate speech present in offensive comments. In this paper we used the first level as it is and a modification of the third level to create a binary hate speech classification task. We used the splits provided in a HuggingFace dataset[7].

5. Experiments[8]

In order to evaluate the hypothesis that tokenizers capable of producing a token sequence better aligned with a morpheme sequence can improve LLMs performance on Brazilian Portuguese, we split our experiment in two steps. First, in Section 5.1, we train new tokenizers and compare them using MorphEval-PT. In the second step, Section 5.2, the tokenizers are used to pre-train new LLMs for Brazilian Portuguese and their performance are compared on downstream tasks, relating the performance on both steps.

[6]Available at: https://huggingface.co/datasets/carolina-c4ai/corpus-carolina

[7]Available at: https://huggingface.co/datasets/ruanchaves/hatebr

[8]The code used to run the experiments are available at:
https://github.com/guilhermelmello/msc_codes

5.1. Evaluating Morphological Coherence in Tokenizers

Tokenizers transform the input text into a sequence of tokens, usually, following the same pipeline: (a) normalization; (b) pre-tokenizer; (c) model; and (d) post-processing. To properly compare Unigram against BPE, we only changed the model in this pipeline. We used NFC Unicode Normalizer, which showed better results in preliminary experiments. The white space were used as pre-tokenizer. We also applied byte-level on both models to eliminate the presence of unknown tokens. In the post-processing step we introduced special tokens (`[EOS]`, `[SEP]` and `[PAD]`). We trained Unigram and BPE tokenizers using the Corpus Carolina (Section 4.1) with vocabulary sizes 5k, 8k, 10k, 30k and 50k. Tokenizers are public available on HuggingFace Hub[9].

The results for MorphEval-PT are divided by morphological process (Tables 3 and 4). In both cases, when evaluating the single token values, which represents the storage route, BPE includes more words in its vocabulary as size increases, an expected behavior of its bottom-up strategy. Even though Unigram show a similar behavior, the growth rate is considerably lower, which shows that Unigram has a preference for using the decomposition route. In our experiments BPE achieved higher precision score than Unigram, in special for inflectional morphemes with gains ranging between 10-30% over Unigram. It is worth noting that this gain is not a result of more segmentations as a single token by BPE. In Table 3, we can observe that BPE required a vocabulary of 5k to obtain a precision of 61.77% with only 0.77% of single tokens. Meanwhile, Unigram required a vocabulary of 50k to obtain a similar precision (61.07%) with 5.29% single tokens. The recall analysis require a careful inspection, as it may represent a better compression rate when associated with a high precision, which can be observed for BPE in Table 3. The opposite, high recall with low precision, is an evidence that tokens are being created unnecessarily, which can be observed for Unigram in Table 4. We also evaluate the presence of morphemes in the vocabulary of each tokenizer. BPE was also able to include more affixes in theirs vocabularies, with 20% to 30% more affixes than Unigram. This difference remained stable when increasing the vocabulary size.

When considering other models in the literature, Albertina [Santos et al. 2024] obtained the better precision and F1 scores for derivational morphemes (Table 3). This is an impressive result when considering that even though it is a Portuguese model, it used the original vocabulary from DeBERTa [He et al. 2021], which was trained on English data. For inflection data (Table 4), Qwen3 [Yang et al. 2025] also obtained impressive results for a multilingual LLM.

5.2. Impact of Morphological Coherence on Downstream Tasks

To evaluate the impact of better morphological coherent tokenizers, we trained LLMs from scratch using Unigram and BPE, called Qwen-PT-Unigram and Qwen-PT-BPE, respectively. We used the 8k vocabulary tokenizers, which show a good precision with a compact vocabulary (Section 5.1). We trained a Qwen3 [Yang et al. 2025] architecture with Causal Language Modeling (CLM) with sequences of 1024 tokens for 10 epochs on Corpus Carolina (Section 4.1). Therefore, the only difference between them is the tokenizer's model. Using batches of 1k, sentences were merged using `[SEP]` in

[9]Available at: https://hf.co/collections/guilhermelmello/tokenizers-qwen-pt

Model	Vocab		Precision	Recall	F1 Score	Single Token
BPE		5k	61.77%	45.90%	52.67%	89 (0.77%)
		8k	63.95%	43.34%	51.66%	224 (1.95%)
		10k	64.20%	42.33%	51.02%	333 (2.90%)
		15k	66.65%	41.63%	51.25%	526 (4.57%)
		30k	70.34%	41.49%	52.20%	1101 (9.57%)
		50k	75.73%	43.07%	**54.91%**	1792 (15.58%)
Unigram		5k	47.03%	68.32%	55.71%	135 (1.17%)
		8k	57.53%	66.59%	**61.73%**	195 (1.70%)
		10k	57.94%	60.86%	59.36%	225 (1.96%)
		15k	59.19%	57.62%	58.39%	291 (2.53%)
		30k	59.21%	53.59%	56.26%	465 (4.04%)
		50k	61.07%	49.80%	54.87%	608 (5.29%)
Qwen3 0.6B	BPE	150k	64.49%	46.04%	53.73%	21 (0.18%)
Albertina 900m	Unigram	128k	74.22%	55.20%	**63.32%**	106 (0.92%)
Albertina 100m	BPE	50k	46.02%	51.53%	48.62%	11 (0.10%)
PeLLE	BPE	50k	51.24%	43.27%	46.92%	68 (0.59%)
BERTimbau	WordPiece	30k	71.16%	46.89%	56.53%	861 (7.49%)
mBERT-cased	WordPiece	120k	62.72%	47.04%	53.76%	326 (2.83%)
TeenyTinyLlama	BPE	32k	70.74%	42.41%	53.03%	940 (8.17%)

Table 3. MorphEval-PT results on derivational dataset measured by precision, recall and F1 Score. Single Token represents the percentage of words that were not splitted in sub-words in a total of 11501 words, which represents the storage route.

Model	Vocab		Precision	Recall	F1 Score	Single Token
BPE		5k	51.44%	58.26%	54.64%	590 (0.20%)
		8k	53.94%	56.20%	55.04%	1207 (0.40%)
		10k	54.98%	55.38%	55.18%	1677 (0.56%)
		15k	57.25%	53.32%	55.21%	2804 (0.94%)
		30k	58.61%	51.11%	54.60%	6467 (2.16%)
		50k	62.47%	49.89%	**55.48%**	11094 (3.70%)
Unigram		5k	30.00%	52.63%	**38.22%**	926 (0.31%)
		8k	29.16%	48.36%	36.38%	1451 (0.48%)
		10k	29.25%	46.67%	35.96%	1796 (0.60%)
		15k	29.16%	44.02%	35.08%	2545 (0.85%)
		30k	30.60%	40.08%	34.71%	3792 (1.26%)
		50k	27.90%	35.00%	31.05%	5204 (1.74%)
Qwen3 0.6B	BPE	150k	58.28%	64.47%	**61.22%**	1020 (0.34%)
Albertina 900m	Unigram	128k	37.30%	43.12%	40.00%	1859 (0.62%)
Albertina 100m	BPE	50k	49.01%	65.03%	55.90%	526 (0.18%)
PeLLE	BPE	50k	26.36%	22.59%	24.33%	1903 (0.63%)
BERTimbau	WordPiece	30k	29.14%	28.22%	28.67%	5765 (1.92%)
mBERT-cased	WordPiece	120k	26.29%	25.10%	25.68%	3277 (1.09%)
TeenyTinyLlama	BPE	32k	60.76%	56.78%	58.70%	5819 (1.94%)

Table 4. MorphEval-PT results on inflectional dataset measured by precision, recall and F1 Score. Single Token represents the percentage of words that were not splitted in sub-words in a total of 299807 words, which represents the storage route.

	ASSIN		ASSIN 2		HateBR		
Model	RTE	STS	RTE	STS	Offensive	Hate	*avg*
Qwen-PT-Unigram	0.6826	0.7377	0.8537	0.7187	0.8999	0.8585	*0.7918*
Qwen-PT-BPE	0.7309	0.7454	0.8589	0.7635	0.9157	0.8871	*0.8169*
Albertina-100m	0.8081	0.8229	0.8792	0.7810	0.8905	0.8428	*0.8374*
BERTimbau-base	0.8172	0.7905	0.8991	0.7903	0.9142	0.8627	*0.8456*
BERTimbau-large	0.8254	0.6616	0.8941	0.7995	0.9249	0.8596	*0.8275*
TeenyTinyLlama-160m	0.6998	0.7110	0.8723	0.6119	0.9071	0.8142	*0.7693*
TeenyTinyLlama-460m	0.7279	0.6810	0.8799	0.6956	0.9192	0.8683	*0.7953*

Table 5. Results on ASSIN, ASSIN 2 and HateBR tasks. RTE and HateBR tasks is measured by F1 Score and STS tasks is measured by Pearson Correlation.

between and splitted in chunks of 1024 tokens. `[PAD]` tokens were inserted on left side when needed, preceded by `[EOS]` token. In preliminary experiments, smaller models obtained better performance and we used L=12, H=1024, A=16 to obtain 0.2B models.

Since it is a decoder architecture, the last token representation was used to fine-tune on ASSIN, ASSIN 2 and HateBR tasks. For each downstream task, we performed a grid search with batch size (4, 8, 16 and 32) and learning rate (5e-3, 5e-4, 5e-5 and 5e-6) for 5 epochs with 42 as a random seed. With selected hyperparameters, the fine-tuning restarted from scratch with random classification weights for 10 epochs, always selecting the best training model. Results show us that BPE consistently outperformed Unigram with an average gain of +2.51 (Table 5). In general, models based on encoder architecture obtained better results. When considering a decoder architecture, Qwen-PT models obtained better results than TeenyTinyLlama models on both ASSIN, ASSIN 2 and HateBR tasks. Also, as expected, bigger models obtained better performance.

6. Conclusions

Motivated by psycholinguistics theories of morphological processing, this paper explored the hypothesis that a sequence of tokens better aligned as a sequence of morphemes can improve LLMs performance on Brazilian Portuguese. We presented MorphEval-PT, an objective evaluation method that helps to assess how likely tokenizers can generate tokens that resemble morphemes. The promising results show evidences that collaborate with literature, but more experiments are required to confirm this hypothesis.

Even though the literature is not consistent on which tokenization method is more efficient [Bostrom and Durrett 2020, Kudo 2018], BPE consistently outperformed Unigram in our experiments, providing evidences that a morphological coherence is a good feature to assess LLM performance on downstream tasks (Section 5.2). By using MorphEval-PT metrics we could observe that BPE contained more morphemes in its vocabulary and was able to create tokens that better align with a morphological coherent segmentation (Section 5.1). Therefore, MorphEval-PT is an important tool that can be used to help the development of new LLMs for Brazilian Portuguese. By providing metrics that can be easily accessed prior any pretraining, MorphEval-PT can reduce the cost related to LLM development by assisting in iterative improvement of the tokenizer. To overcome its limitations, as future work, a more robust dataset needs to be developed.

7. Acknowledgments

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP 2019/07665-4) and by the IBM Corporation.

8. Limitations

As already discussed in Section 3, the proposed evaluation method MorphEval-PT contains a series of limitation inhered from MorphyNet. These limitation were addressed by the interpretation of evaluation metrics (Section 3.3). Therefore, MorphEval-PT evaluation procedure cannot state an overall capacity of tokenizers in generating a morpheme segmentation. While these limitation are not overcome, MorphEval-PT can only support evidences of the ability of generating derivational and inflectional morphemes.

As stated by [Hou et al. 2023], the gain in performance that may result by the presence of morphemes may not be statistically significant. To overcome this limitations, more experiments must be executed to support and validate if the results in this paper are significant and hold with different training settings, with different LLM architectures and vocabulary sizes. Also, other tokenizers can be considered, like Morfessor [Creutz and Lagus 2002] and StateMorph [Nouri and Yangarber 2017].

9. Disclosure on Generative AI Usage

During the development of this work, we used ChatGPT to assist with experiment's code generation. Every generated line of code was manually validated and edited by the authors. During the writing of this paper, no generative AI was used and the authors take full responsibility for the content of this publication.

References

Batsuren, K., Bella, G., and Giunchiglia, F. (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. Association for Computational Linguistics.

Beesley, K. R. and Karttunen, L. (2003). *Finite-state morphology*. CSLI Publications.

Bostrom, K. and Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

Crespo, M. C. R. M., de Souza Jeannine Rocha, M. L., Sturzeneker, M. L., Serras, F. R., de Mello, G. L., Costa, A. S., Palma, M. F., Mesquita, R. M., de Paula Guets, R., da Silva, M. M., Finger, M., de Sousa, M. C. P., Namiuti, C., and do Monte, V. M. (2023). Carolina: a general corpus of contemporary brazilian portuguese with provenance, typology and versioning information.

Creutz, M. and Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Creutz, M. and Linden, B. (2004). Morpheme segmentation gold standards for finnish and english.

de Alencar, L. F., Cuconato, B., and Rademaker, A. (2018). Morphobr: An open source large-coverage full-form lexicon for morphological analysis of portuguese. *Texto Livre: Linguagem e Tecnologia*, 11(3):1–25.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fonseca, E. R., dos Santos, L. B., Criscuolo, M., and Aluísio, S. M. (2016). Visao geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13.

Gonçalves, C. A. (2019). *Morfologia*. Parábola, 1st edition.

He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Hofmann, V., Pierrehumbert, J., and Schütze, H. (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. Association for Computational Linguistics.

Hou, J., Katinskaia, A., Vu, A.-D., and Yangarber, R. (2023). Effects of sub-word segmentation on performance of transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7413–7425. Association for Computational Linguistics.

Kudo, T. (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. Association for Computational Linguistics.

Leminen, A., Smolka, E., Duñabeitia, J. A., and Pliatsikas, C. (2019). Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *Cortex*, 116:4–44.

Li, X., Meng, Y., Sun, X., Han, Q., Yuan, A., and Li, J. (2019). Is word segmentation necessary for deep learning of Chinese representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Mota, C., Carvalho, P., and Barreiro, A. (2016). Port4NooJ v3.0: Integrated linguistic resources for Portuguese NLP. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1264–1269. European Language Resources Association (ELRA).

Nouri, J. and Yangarber, R. (2016). A novel evaluation method for morphological segmentation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3102–3109. European Language Resources Association (ELRA).

Nouri, J. and Yangarber, R. (2017). Learning morphology of natural language as a finite-state grammar. In *Statistical Language and Speech Processing*, pages 44–57, Cham. Springer International Publishing.

Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., and Schwartz, L. (2021). Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Plag, I. (2018). *Word-formation in English*. Cambridge university press.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Real, L., Fonseca, E., and Gonçalo Oliveira, H. (2020). The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

Santos, R., Rodrigues, J., Gomes, L., Silva, J., Branco, A., Cardoso, H. L., Osório, T. F., and Leite, B. (2024). Fostering the ecosystem of open neural encoders for portuguese with albertina pt-* family.

Schuster, M. and Nakajima, K. (2012). Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Sennrich, R., Haddow, B., and Birch, A. (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. Association for Computational Linguistics.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. (2025). Qwen3 technical report.