Exploring morphology-aware tokenization: A case study on Spanish language modeling

Alba Táboas García[1] **and Piotr Przybyła**[1,2] **and Leo Wanner**[3,4]
[1] TALN Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[3] Barcelona Supercomputing Center (BSC)
[4] Catalan Institute for Research and Advanced Studies (ICREA)
alba.taboas@upf.edu, piotr.przybyla@upf.edu, leo.wanner@bsc.es

Abstract

This paper investigates to what extent the integration of morphological information can improve subword tokenization and thus also language modeling performance. We focus on Spanish, a language with fusional morphology, where subword segmentation can benefit from linguistic structure. Instead of relying on purely data-driven strategies like Byte Pair Encoding (BPE), we explore a linguistically grounded approach: training a tokenizer on morphologically segmented data. To do so, we develop a semi-supervised segmentation model for Spanish, building gold-standard datasets to guide and evaluate it. We then use this tokenizer to pre-train a masked language model and assess its performance on several downstream tasks. Our results show improvements over a baseline with a standard tokenizer, supporting our hypothesis that morphology-aware tokenization offers a viable and principled alternative for improving language modeling.

1 Introduction

The way tokenization is performed has been shown to be essential for the performance of neural language models. Early embeddings such as WORD2VEC (Mikolov et al., 2013) and GLOVE (Pennington et al., 2014) treated whole words as tokens, producing semantically meaningful representations, but at the cost of large vocabulary sizes and sensitivity to out-of-vocabulary words. To address these limitations, i.e., to reduce vocabulary size and improve generalization, research shifted to subword tokenization, which splits a continuous sequence of characters into frequent character subsequences. Over time, the use of subword tokenizers such as SentencePiece (Kudo and Richardson, 2018), WordPiece (Wu et al., 2016), Unigram (Kudo, 2018), or Byte Pair Encoding (BPE) (Sennrich et al., 2016) has become a must in state-of-the-art NLP models. In particular, BPE established itself as a *de facto* standard. Still, despite their advantages, these tokenizers reveal an important drawback: they operate on purely statistical patterns in the data, which optimizes for frequency and compression, but is agnostic to the internal structure of the words. As a consequence, the resulting subword tokens are frequently misaligned with morphological boundaries (Church, 2020). This misalignment has been shown to impact the models' ability to represent and generalize morphological information, ultimately limiting the performance of the models in downstream tasks that rely on fine-grained linguistic cues (Hofmann et al., 2020, 2021; Klein and Tsarfaty, 2020; Bostrom and Durrett, 2020; Tan et al., 2020). Nonetheless the picture is not settled: other recent studies report only marginal gains or even advantages for purely statistical approaches (Saleva and Lignos, 2021; Truong et al., 2024; Arnett et al., 2024; Arnett and Bergen, 2025). Against this backdrop, a growing body of work explores linguistically-informed tokenization. Jabbar (2024) does so for English, Toraman et al. (2023) for Turkish, Park et al. (2020) for Korean and Westhelle et al. (2022) for Brazilian Portuguese. While Jabbar (2024) demonstrates gains by integrating morphology into subword tokenization, their strategy can hardly be applied to morphologically richer languages since it relies on static vocabularies that struggle with unseen forms and linguistic variability. Results from the other studies are less clear-cut: some linguistically informed models perform better but often require rule-based preprocessing, while in other cases linguistically agnostic models outperform them.

Therefore, we investigate to what extent morphologically informed tokenization can improve linguistic modeling and downstream task performance in Spanish. This language has so far remained underexplored in this context, despite the challenges posed by the fusional nature of its morphology, where multiple grammatical features are

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30505–30518
November 4-9, 2025

encoded in a single morpheme. To ensure portability of the proposed approach, we infuse morphological knowledge at the training stage of the tokenizer, without that any changes to the tokenization algorithm or the language model architecture are required. Our strategy results in a fully integrated tokenizer that avoids extra preprocessing, as in the rule-based approach of Toraman et al. (2023), nor relies on static vocabularies with complex detokenization schemes, as Jabbar (2024) does.

First, we develop a segmentation model for Spanish using MorphAGram (Eskander et al., 2020), which shows a considerably higher quality in morphological segmentation compared to Morfessor (Smit et al., 2014; Grönroos et al., 2014) used, e.g., by Westhelle et al. (2022). The segmentation model is applied to a dataset to obtain linguistically meaningful sub-word units, on which we train a standard BPE tokenizer. To evaluate the linguistic quality of both the segmentation model and the resulting tokenizer, we contrast their outputs against two manually annotated word lists totaling over 7,000 entries that we created for this purpose.

Finally, we pretrain a RoBERTa-based language model with the resulting morphology-aware tokenizer and a novel left-to-right within-word masking strategy inspired by recent evaluation practices (Kauf and Ivanova, 2023), and compare it with a baseline model trained with a standard BPE. Our evaluation combines intrinsic and extrinsic metrics, including perplexity, word prediction accuracy, and performance on common downstream tasks such as natural language inference, paraphrase detection, and semantic text similarity, as well as a more fine-grained linguistic evaluation on the model's morpho-syntactic capabilities.

Our experiments confirm that incorporating morphological awareness into the tokenization process consistently enhances language modeling performance for Spanish over a range of different tasks, reinforcing the idea that morphology-informed tokenizers provide a robust and linguistically grounded alternative to purely statistical standard approaches.

2 Related work

Most of the state-of-the-art LM applications use BPE (Sennrich et al., 2016) for subword tokenization with little scrutiny. However, a growing body of research highlights limitations of purely statistical pattern-driven subword tokenization. For instance, Hofmann et al. (2021) study how tokenization affects the way models internalize complex morphology, focusing on BERT representations of derivationally complex words. For languages with rich morphology, the inadequacy of purely statistical subword units becomes even more apparent; see, e.g., Klein and Tsarfaty (2020), who probe how well BERT captures morphological information in Hebrew and find that its default word pieces fail to reflect meaningful morphemes.

While statistical tokenizers remain practical and performant, their linguistic blind spots can hinder model efficiency, generalization, and downstream interpretability, as has been shown by a wave of work on morphologically informed alternatives. Thus, Bostrom and Durrett (2020) show that Unigram (Kudo, 2018) produces subwords that are better aligned with morphological boundaries and yields equal or superior performance on downstream tasks than BPE. Cross-linguistic evidence also points to the limits of BPE: Park et al. (2021) show that BPE tokenization fails to adequately capture morphological structure across languages, with surprisal strongly correlated to morphological complexity, while morphology-aware methods yield more robust language modeling performance. Other studies reinforce this trend: Tan et al. (2020) shows that adding morphological information to the tokenization stage increases language modeling robustness to inflectional variation in L2 and World Englishes, and Bauwens and Delobelle (2024) prune BPE's vocabulary with morphological semi-supervision making it better aligned to derivational and compound boundaries in English, Dutch and German, and find improvements in downstream tasks for Dutch.

Building on these insights, a number of works explicitly implement morphology-aware tokenization schemes and analyze their effect on downstream performance. MorphPiece (Jabbar, 2024) demonstrates performance gains by integrating morphological segmentation into subword tokenization, but it is limited to English and relies on static vocabularies that struggle with unseen forms and linguistic variability. Other work has explored linguistically-motivated tokenization for morphologically-richer languages. Thus, in Turkish, Toraman et al. (2023)'s morphologically-informed tokenizer rivals much larger models despite a simpler architecture, but implies preprocessing by a rule-based morphological analyzer. Similar patterns have been observed for Korean (Park et al., 2020), where a hybrid tokenizer achieves

strong downstream results combining morphology and subword units. In Brazilian Portuguese, morphological tokenization based on Morfessor (Smit et al., 2014) has shown advantages over standard WordPiece (Westhelle et al., 2022).

At the same time, the research landscape is more nuanced, and several studies report that statistical tokenization can outperform morphological strategies or that the latter provide only marginal gains (e.g., Saleva and Lignos, 2021; Zhu et al., 2019; Banerjee and Bhattacharyya, 2018). These findings, however, need to be understood with care. For example, Arnett and Bergen (2025) argue that morphological alignment does not explain performance differences, but their comparisons are across languages with different morphological typologies (fusional vs. agglutinative), rather than across tokenization strategies within the same language, which is the focus of our work. Similarly, some studies that report no benefits from morphology-aware segmentation examine only narrow linguistic phenomena. For instance, Truong et al. (2024) restrict their evaluation to affixal negation in English, and Arnett et al. (2024) analyze only the plural forms of Spanish nouns. Such focused tests do not necessarily reflect the broader impact that morphology-aware tokenization can have on the intrinsic quality of language models or on their downstream applications.

Our work contributes to this ongoing discussion by focusing on Spanish, which has received so far little attention from the perspective of subword tokenization. In contrast to, e.g., the agglutinative morphology of Turkish and Korean, its morphology is fusional, as in Portuguese, and can thus be expected to be more challenging to capture. In contrast to the previous works such as (Jabbar, 2024) on English or (Toraman et al., 2023) on Turkish, our tokenzation strategy is not limited to predefined lists, does not involve convolute detokenization strategies, and does not imply any preprocessing stages. Also, in contrast to (Westhelle et al., 2022)'s study on Portuguese, we rely on MorphAGram (Eskander et al., 2020) instead of Morfessor, as our preliminary experiments (see Section 4.1) showed that it produces much higher-quality morphological segmentations.

3 Morphology-Aware Tokenization

Morphological segmentation and subword tokenization share a common objective: breaking down words into smaller units, but they differ fundamentally in their guiding principles. While the former aims for linguistically meaningful components (*morphs*)[1], the latter relies only on co-occurrence statistics to identify frequent character sequences. For example, the word *undeniability* may be morphologically segmented as *un+ deni + abil + ity*[2], whereas a standard tokenizer like RoBERTa's BPE splits it into *unden + iability*.

We combine morphological segmentation and tokenization in a two-stage approach. First, we apply segmentation to generate linguistically-informed subword units, and then we use these units to train the tokenizer.

3.1 Morphological segmentation for Spanish

For segmentation, we adapt MorphAGram (Eskander et al., 2016, 2018, 2020), which proved to consistently outperform or match the performance of other morphological segmenters on a wide range of morphologically diverse languages (Eskander et al., 2020, 2021, 2022; Tan Le et al., 2022; Okabe and Yvon, 2023). MorphAGram is based on *Adaptor grammars*, i.e., non-parametric Bayesian extensions of Probabilistic Context-Free Grammars (PCFGs) (Johnson et al., 2007). MorphAGram models learn a PCFG and a collection of frequent morphs directly from an input lexicon (a supplied list of words) and an initial grammar which specifies the internal structure of words. The framework provides several built-in grammars; we use the one that performed best in our experiments. Figure 1 illustrates its output for the word *irreplaceables*.

While an input lexicon and a PCFG are sufficient to provide output of reasonable quality, MorphAGram also supports a semi-supervised learning setup where prior linguistic knowledge can be seeded within the PCFG as a list of known affixes.

To adapt MorphAGram to Spanish, we compiled a 50,000-word Spanish lexicon by combining two sources: the list of words and their frequencies (derived from real-word usage) in the CREA corpus[3] and the complete *Real Academia Española* (RAE) dictionary. Words not listed in the RAE dictionary have been filtered from the corpus. On the

[1]Morphological segmentation can be either canonical —*morpheme*-based and thus, better aligned with morphological theory— or surface —*morph*-based, which is more practical from a tokenization standpoint, since it allows for easy reconstruction of the original word. Throughout our study we will always use surface segmentation.

[2]See the slight difference with the corresponding canonical segmentation: *un + deny + **able** + ity*

[3]Corpus de Referencia del Español Actual https://corpus.rae.es/lfrecuencias.html

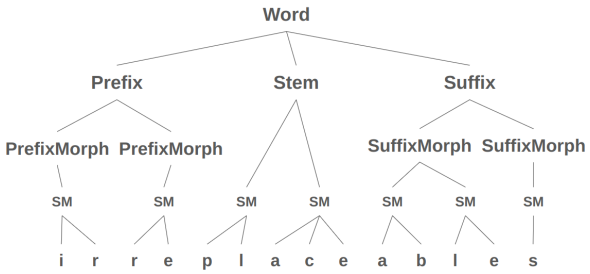

Figure 1: Segmentation of the English word *irreplaceables* using the selected grammar.

other side we include plurals, clitic-attached verb forms, feminine participles, diminutives, superlatives, and also adverbs ending in *-mente* (such as, e.g., *claramente* 'clearly'), which are represented in RAE through their base form only. For the semi-supervised setup, we compiled a list of affixes, including verb conjugation paradigms, known gender and number suffixes, and derivational affixes from RAE's compendium[4]. In addition to the affixes used during training, we also incorporated into the model a list of common invariable words (including prepositions, conjunctions, and frequent adverbs and interjections) that should never be segmented.

3.2 Tokenization for Spanish

To incorporate morphological knowledge while remaining compatible with state-of-the-art subword tokenization, we developed a hybrid tokenizer that combines linguistically-motivated and statistical subword segmentations. To this end, we apply the segmentation model presented in Subsection 3.1 to a dataset so that each word is previously segmented into morphs before training a BPE tokenizer (Sennrich et al., 2016) on this pre-segmented dataset[5].

We opted for this approach instead of using the segmentation model directly as a tokenizer for several reasons. First, while the morphological segmentation model can generalize to unseen words, performing inductive segmentation at runtime for every word would be computationally inefficient, particularly in large-scale applications. Second, although the segmentation model generalizes beyond its training data, it does so imperfectly. For instance, it relies on a finite list of compatible prefix–suffix patterns and sometimes fails to segment novel words that fall outside these learned combinations, defaulting to leaving them unsegmented. Also, while the segmentation model may identify linguistically valid but exceedingly rare morphs, including them in a tokenizer's vocabulary would be inefficient for language modeling purposes. By training BPE on top of morphologically segmented data, we strike a balance between linguistic informativeness and statistical efficiency. Finaly, integrating the BPE tokenizer into existing language modeling pipelines is more straightforward using tools such as the HuggingFace Transformers library (Wolf et al., 2020), which expects standard tokenizer interfaces.

4 Subword tokenization experiments

Prior to the analysis of the influence of morphologically-informed tokenization on downstream tasks, we carried experiments to assess the quality of morphologically-informed segmentation itself and its effects on tokenization.

4.1 Morphological segmentation experiments

We trained a fully unsupervised MorphAGram model and a semi-supervised model as described in Section 3.1, with the same parameters[6] as recommended in the original work (Eskander et al., 2020). Training took around 5.5h each on a single HPC node (24-core Xeon node with 96GB RAM), with limited parallel efficiency. Table 1 illustrates typical outputs of both models, contrasting them with the output of a Morfessor 2.0 baseline (Virpioja, 2013). We used the Morfessor off-the-shelf Spanish model from Polyglot[7] as a simple and widely used baseline, rather than a semi-supervised setup, since our aim was merely to provide a common point of comparison; prior work (Eskander et al., 2020, 2021, among others) has already shown adaptor grammars to outperform Morfessor. We also provide gold segmentations to highlight the models' limitations. The results demonstrate that the segmentations provided by the semi-supervised MorphAGram model are, overall, more closely aligned with the references, especially considering what they identify as stems for the words.

[4] https://www.rae.es/sites/default/files/Elementos_compositivos_prefijos_y_sufijos_del_espanol_Esencial.pdf

[5] Note that this preprocessing step is required only once: the segmentation model is used during tokenizer training to preprocess the data. Afterward, both the tokenizer and the language model are applied directly to raw text.

[6] See https://github.com/rnd2110/MorphAGram/tree/0ccf074149baf78735c0f5adcc359a0f90e96f35.

[7] Via https://github.com/aboSamoor/polyglot/tree/master. Unfortunately, the training details for this model are not disclosed.

Morfessor 2.0	MorphAGram Unsupervised	MorphAGram Semi-supervised	Reference
impre-visible	impre-vis-**ible**	im-pre-**vis**-ible	im-pre-**vis**-ible
inter-nacional	inter-n-**acional**	**intern**-a-cion-al	inter-**nacion**-al
des-cuida-da-mente	des-cuid-**adamente**	des-**cuid**-ada-mente	des-**cuid**-ada-mente
rápida-mente	rá-pid-**amente**	**ráp**-ida-mente	**rápid**-a-mente
transforma-ción	trans-form-**ación**	trans-**form**-ación	trans-**form**-ación
re-conocimiento	re-conoc-**imiento**	**reconoc**-imiento	re-**conoc**-imiento
re-formula-mos	re-formul-**amos**	re-**formul**-amos	re-**formul**-amos
configura-s-te	con-figur-**aste**	**configuraste**	con-**figur**-aste

Table 1: Segmentation examples from the adapted MorphAGram models, Morfessor 2.0, along with reference segmentations (stem morphs in bold). Note that Morfessor models do not differentiate between stems and affixes.

Morphology-aware
La ␣hero ica ␣ciudad ␣dorm ía ␣la ␣siest a . ␣El ␣v iento ␣Sur , ␣cal iente ␣y ␣perez oso , ␣empuj aba ␣las ␣nub es ␣blanque cin as ␣que ␣se ␣rasg aban ␣al ␣corr er ␣hacia ␣el ␣Norte .
Standard BPE
La ␣heroica ␣ciudad ␣dormía ␣la ␣siesta . ␣El ␣viento ␣Sur , ␣caliente ␣y ␣perez oso , ␣empu jaba ␣las ␣nubes ␣blanque cinas ␣que ␣se ␣ras gaban ␣al ␣correr ␣hacia ␣el ␣Norte .

'The heroic city was taking a nap. The hot, lazy South wind pushed the chalky clouds, which tore as they raced North.'

Table 2: Tokenization samples from the morphologically informed tokenizer and the standard BPE. Blanks mark the separation between tokens; '␣' stands for the beginning of a word; differing tokenizations appear in color: teal for the morphology-aware strategy, purple for BPE.

To obtain a more objective picture, we evaluated both the unsupervised and the semi-supervised configurations of our MorphAGram adaptation. Since we could not find publicly available reference segmentations for Spanish[8], we manually curated our own. We selected 1,200 unique words from the CREA corpus ensuring none of them overlapped with the lexicon used for training. Additionally, to assess the model's performance on naturalistic input, we annotated approximately 5,400 words drawn from short texts randomly sampled from the Spanish portion of the AnCora Universal Dependencies dataset (Taulé et al., 2008). Our segmentations were produced following linguistic criteria grounded in RAE guidelines, drawing on Spanish compositional elements and affixes, as well as the most recent edition of the official Spanish grammar (Real Academia Española, 2010). Note that we did not aim for maximal morphological granularity. In particular, gender and number markers in nouns, as well as tense, aspect, mood, person, and number in verbs were not separated.

Table 3 compares the segmentation accuracy of both MorphAGram configurations against the Morfessor 2.0 baseline. We report the F1 score for two different metrics: Boundary Precision and Recall (BPR) and EMMA-2 (Virpioja et al., 2011)[9] on word-level and text-level segmentations, with reference to our manually annotated data. BPR is the traditional metric for evaluating morphological segmentation, assessing how well the predicted boundaries align with reference segmentation. In contrast, EMMA-2 shifts the focus to morph-level matching, allowing multiple predicted morphs to map to a single reference one.

The semi-supervised MorphAGram model (AGSS) achieves the highest scores across all settings, significantly outperforming both the baseline and the unsupervised variant (AGUS). This confirms the effectiveness of incorporating linguistic knowledge into the segmentation process, and supports our decision to use this specific model for the following steps.

4.2 Morphological tokenization experiments

We took a random 10% subset of the Spanish portion of the OSCAR corpus as our tokenizers' training dataset. We applied the semi-supervised Mor-

[8]Canonical segmentations were available, but not surface segmentations.

[9]As implemented for MorphoChallenge shared tasks: http://morpho.aalto.fi/events/morphochallenge/

Segmentation Model	Words		Texts	
	BPR	EMMA2	BPR	EMMA2
Morfessor	0.29	0.72	0.64	0.74
AGUS	0.68	0.78	0.84	0.84
AGSS	**0.77**	**0.88**	**0.89**	**0.89**

Table 3: Segmentation model evaluation. Morfessor 2.0 baseline against our unsupervised (AGUS) and semi-supervised (AGSS) MorphAGram models.

phAGram model to segment the dataset so that each word in the subset is segmented into morphs, and a special boundary symbol (which we include as a special token in the tokenizer's specifications) is inserted between morphs. Segmentation took 12.4h on an Intel Core i7 (with 8 cores and 16GB RAM). We then trained two BPE tokenizers with a 50K-size vocabulary, using the HuggingFace Transformers library (Wolf et al., 2020) on the same HPC node (24-core Xeon node with 96GB RAM). One was trained on the morphologically pre-segmented dataset and the other on the raw dataset, taking 7.7h and 9.4h respectively (both single-threaded). The remaining training settings (initial alphabet, pre-tokenizer, postprocessor, etc.) were the same for both tokenizers. Table 2 shows the differences in tokenization outputs between our morphology-aware tokenizer and the standard BPE, while Figure 2 illustrates how the subwords produced by the morphological tokenizer better align with our reference segmentations than those from BPE.

We also evaluated both tokenizers using the BPR and EMMA-2 metrics on our two manually annotated gold segmentation datasets to assess their morphological quality, along with their subword fertility (the average number of tokens per word), a more standard evaluation metric. As Table 4 shows, the morphologically informed tokenizer substantially outperformed the standard BPE model in terms of segmentation accuracy, while resulting in a higher subword fertility due to the finer granularity of the morph-based tokens.

5 Language modeling and applications

To explore the impact of morphologically informed tokenization, we conducted a series of experiments incorporating both our morphology-aware tokenizer and the standard BPE baseline into a Spanish language model. We compare the resulting models, first evaluating their core language modeling capabilities, and then assessing their performance on a selection of downstream applications.

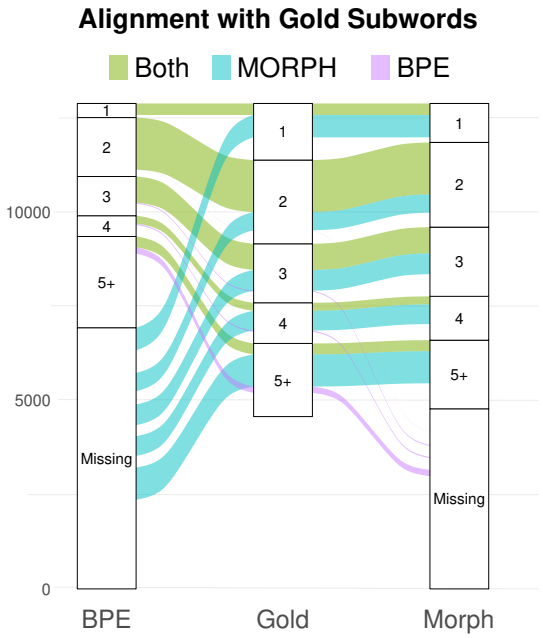

Figure 2: Subword alignment between the morphological or standard BPE and our gold segmentations. The figure shows the number occurrences (Y axis) of subwords of different length (1, 2, ... 5+ characters) from the gold subset that are also present in standard BPE (purple), morphological variant (teal) or both (green).

5.1 Impact on language modeling

For our experiments, we use a standard RoBERTa-based masked language model with 12 transformer layers, 768 hidden dimensions, and 12 attention heads for Spanish. In order to isolate the effect of tokenization, we trained it, on the one hand, with the standard BPE tokenizer, which was trained on raw text data, and, on the other hand, with our morphologically informed tokenizer (Section 3.2).

Both model variants (referred to as 'morphology-aware' and 'vanilla' model, respectively) were pre-trained from scratch on a randomly selected 16GB subset of the FineWeb-2 corpus (Penedo et al., 2024). Inspired by recent recommendations for masked language models (Kauf and Ivanova, 2023), we experimented with three different masking strategies during pre-training: *naive*, *left-to-right* (L2R), and *whole-word* (WW). *Naive* masking is the original dynamic masking strategy for the RoBERTa models: a random 15% of tokens is masked and loss is computed over them. In the L2R setup, 12% of tokens are randomly selected as prediction targets and all subsequent tokens within the same word are masked as well (however, loss is computed only on the originally selected ones). WW masking is similar, but only a 10% of tokens

Tokenizer	Words BPR	Words EMMA2	Texts BPR	Texts EMMA2	Subword fertility	Tokenization examples		
Morph-aware	**0.67**	**0.84**	**0.83**	**0.84**	1.45	␣in comprens ible	␣camin os	␣des activ ase
Standard	0.39	0.74	0.70	0.68	1.12	␣incomprensible	␣cam inos	␣desac tivas e

Table 4: Tokenizer evaluation: subword fertility and morph detection accuracy using BPR and EMMA2 metrics.

are selected for prediction and all other tokens pertaining to the same words are also masked. L2R and WW aim to preserve word-internal coherence during training, which is especially important for morphologically segmented input.

The model is then trained using the HuggingFace Transformers library (Wolf et al., 2020). Apart from the aspects defined above, the training parameters were set as recommended in the `TrainingArguments` class. 5% of the dataset were set aside as a development subset to monitor perplexity. We trained for 5 epochs using a computing node with two NVIDIA H100 cards, which lasted 53 or 91 hours for the baseline and morphology-aware variant, respectively. To assess the models' intrinsic quality, we evaluate their language modeling capabilities using perplexity and word prediction accuracy, and test how well they handle agreement phenomena.

Perplexity measurements

We computed the perplexity of both variants of the model over a Spanish dataset distinct from the pre-training data. For the computation, we used the same three masking strategies employed during pre-training: *naive*, *whole-word*, and *left-to-right*.

As shown in Table 5, the morphology-aware model performs best with the left-to-right within-word masking strategy, while the vanilla model that uses a standard BPE tokenizer achieves a lower BPB with naive masking. Based on these results, we retained only the best-performing configuration for the vanilla model in the subsequent tasks.

Perplexity has long been a standard evaluation metric, but direct comparisons across models with different tokenizers are not completely fair. For this reason, we additionally report a normalized metric such as bits-per-byte (BPB), c.f. Table 6. However, the choice of metric does not affect our conclusions: the best-performing vanilla and morph-aware models remain the same.

LAMBADA word prediction

For the word prediction experiment, we used a machine-translated version of the original English

Tokenizer	Masking	Perplexity Naive	Perplexity L2R	Perplexity WW
Standard BPE	Naive	**10.11**	**14.46**	**20.51**
Standard BPE	L2R	16.88	15.82	21.54
Standard BPE	WW	27.07	24.76	20.61
Morph-aware	L2R	**7.61**	**7.76**	20.71
Morph-aware	WW	23.41	25.64	**20.03**

Table 5: Perplexity of evaluated models under three masking strategies: *naive*, *left-to-right* (L2R), and *whole-word* (WW). Rows show the masking strategy during pre-training, and columns show that applied during evaluation. Lower perplexity values indicate better language modeling quality. Boldface marks lowest perplexity within each tokenizer family: BPE vs. morph-aware.

Tokenizer	Masking	Bits-per-byte Naive	Bits-per-byte L2R	Bits-per-byte WW
Standard BPE	Naive	**0.657**	**0.759**	**0.859**
Standard BPE	L2R	0.803	0.784	0.872
Standard BPE	WW	0.938	0.912	0.860
Morph-aware	L2R	**0.781**	**0.789**	1.168
Morph-aware	WW	1.214	1.249	**1.155**

Table 6: Bits-per-byte of evaluated models under three masking strategies: *naive*, *left-to-right* (L2R), and *whole-word* (WW). BPB measures the average number of bits required to encode each byte of text. Lower values correspond to more efficient modeling of the data. Boldface marks lowest BPB within each tokenizer family: BPE vs. morph-aware.

Task	Model Vanilla	Model Morph-L2R	Model Morph-WW
LAMBADA word prediction (Acc.)			
Final word	0.338	**0.400**	0.370
Random word	0.393	**0.434**	0.381
Morpho-syntactic tests (Acc.)			
Agreement	0.864	**0.926**	0.745

Table 7: Word prediction and agreement tests accuracy of a vanilla baseline and two morphology-aware models, trained with *left-to-right* (L2R) and *whole-word* (WW) masking. Bold marks the best model per task.

dataset (Paperno et al., 2016), which challenges models to predict the final word of a narrative passage. The task is designed to require sentence and discourse-level understanding, since the last word is typically not recoverable from the local context alone. Translation quality is especially critical here, hence, we manually reviewed a few examples to ensure the dataset's adequacy. To reduce the bias introduced by proper nouns, many of which were untranslated English names, we added a second task in which a randomly selected non-final word in each passage was also selected as a prediction target. This facilitated the inclusion of a broader range of parts of speech and syntactic contexts.

As our models are bidirectional rather than generative, and since the target words could consist of multiple tokens (especially for the morphology-aware model with its finer-grained vocabulary), we adopted a greedy prediction strategy. Each word was predicted in a both left-to-right and right-to-left fashion, the final choice being the version with higher joint probability. Accuracy was measured as an exact match between predicted and target words. As shown in Table 7, in this task, the morphology-aware model with *left-to-right* masking achieves an accuracy of 40.0% compared to 37.0% with *whole-word* masking, and 33.8% for the vanilla model when predicting the final word of the texts, and a 43.4% vs. 38.1% and 39.3%, respectively, for the randomly selected word.

Morpho-syntactic agreement tests

To have a more linguistically-oriented assessment of the performance of the models, we also examined their behavior on a set of controlled morpho-syntactic tests targeting agreement phenomena in Spanish. The tests, which reveal the model's grasp on morpho-syntactic dependencies, are drawn from SyntaxGym ES (Pérez-Mayos et al., 2021), which adapts the SyntaxGym methodology (Hu et al., 2020; Gauthier et al., 2020) to Spanish. To this end, the model is presented with two or more nearly identical sentence variants, differing only in the agreement features of a specific target word. Only one variant is grammatically correct (in which the target word agrees in number and gender or number and person with its controller), while the other variants contain mismatches. The model is expected to assign lower *surprisal*, i.e., higher probability, to the grammatically correct target than to any of the incorrect ones. The tests cover a range of contexts: nominal agreement within noun phrases (noun-article, noun-adjective) and at the clause level between subject nouns and predicative attributes and complements, as well as verbal agreement between subject nouns or pronouns and finite verbs. They result in 92.6% accuracy for the morphology-aware model with *left-to-right* masking against 74.5% with *whole-word* masking and 86.4% for the vanilla baseline; cf. Table 7.

Considering the results from both the LAMBADA word prediction task and the agreement tests, the rest of our study, which is computationally more demanding, is conducted only for the morphology-aware model with *left-to-right* masking and the vanilla baseline.

	Model	
Task	Vanilla	Morph-aware
Natural language inference		
XNLI (Accuracy)	0.733	**0.742**
InferES (Accuracy)	0.656	**0.666**
Paraphrase identification		
PAWS-X (F1)	0.841	**0.845**
Semantic text similarity		
STS (Combined)	0.776	**0.801**

Table 8: Downstream performance of a vanilla baseline and the selected morphology-aware model. Bold marks the best model per task.

5.2 Downstream applications

To assess how well the two variants of the model transfer to real-word language understanding tasks, we fine-tuned them on three standard downstream tasks commonly used and featured in benchmarks such as EvalES[10] and GLUES (Canete et al., 2020): natural language inference, paraphrase identification, and semantic text similarity. For this purpose, we performed a basic hyperparameter search using combinations of batch sizes (8, 16), learning rates (10^{-5}, 3×10^{-5}, 5×10^{-5}), and weight decay values (0.1, 0.01). Training was conducted for either 5 or 10 epochs depending on the task, with a fixed warm-up ratio of 0.1. For each task and dataset, the best combination was chosen based on the validation set performance. The specific train/validation/test splits for each task are detailed below. The final results of the best combination on the corresponding test sets are reported in Table 8.

[10] https://benchmark.plantl.bsc.es/

Natural language inference

We first used the Spanish portion of the Cross-Lingual Natural Language Inference corpus (XNLI; Conneau et al. 2018), which contains 400,202 sentence pairs annotated for entailment, contradiction, or neutrality. 2,490 of these sentence pairs constitute the validation set and other 5,010 pairs constitute the test set. However, while XNLI is a widely adopted benchmark for Spanish LM evaluation, it is based on machine translation of inconsistent quality, which often introduces artifacts that can negatively influence learning. To mitigate this limitation, we also fine-tuned both variants of the model on InferES (Kovatchev and Taulé, 2022) – a smaller, but original Spanish NLI corpus of 8,055 sentence pairs. InferES was specifically designed to be challenging and linguistically rich. It includes contrastive and adversarial examples that target complex linguistic phenomena, such as negation and co-reference. Since the dataset does not contain a validation set, we created one by splitting the original test set into 645 examples for validation and 967 for final testing. The morphology-aware model scores 74.2% accuracy on XNLI and 66.6% on InferES, a point over the baseline in both cases.

Paraphrase identification

For paraphrase identification, we used the Spanish portion of PAWS-X (Yang et al., 2019), a multilingual dataset containing adversarially constructed paraphrase pairs. The dataset includes 49,401 training pairs, 2,000 pairs for the development and 2,000 pairs for the test set. PAWS-X emphasizes lexical overlap while varying syntactic structure, making it especially useful for assessing a model's deeper understanding of sentence semantics. In this particular task, the differences are quite small, with F1 scores going from an 84.1% for the vanilla model to 84.5% for the morphology-aware one.

Semantic text similarity

For this task, we used the STS dataset included in the EvalES benchmark. This dataset was built from the Spanish test sets of SemEval-2014 and SemEval-2015 (Agirre et al., 2014, 2015) and consists of 1,321 sentence pairs for training, 78 for development, and 156 for testing. The task involves predicting a graded similarity score, typically using regression. In this task, the morphology-aware model achieves an accuracy of 80.1%, compared to 77.6% of the vanilla baseline.

6 Discussion and concluding remarks

Our findings show that morphology-aware tokenization provides consistent and meaningful improvements in language modeling for Spanish.

We began by demonstrating that morphological segmentation benefits from the incorporation of linguistic knowledge. Our semi-supervised MorphAGram model outperforms both the unsupervised variant and the Morfessor baseline, supporting its use as the foundation for our tokenizer. Applied at the training stage, this segmentation strategy yields units that better reflect true morphological structure, as evidenced by their stronger alignment with gold-standard segmentations. While this morphologically informed tokenizer introduces higher subword fertility due to its finer granularity, the resulting tokens are more linguistically meaningful and better capture internal word structure.

These gains translate into an overall higher LM quality. Our morphology-aware model consistently outperforms the baseline on all evaluated tasks. The strongest gains are seen in morphologically sensitive settings, with a 6-point improvement in the word prediction task and a 7-point accuracy boost in the morpho-syntactic agreement tests. General-purpose tasks that do not explicitly target morphology (e.g., natural language inference, paraphrase detection and semantic text similarity) also reflect consistent, albeit smaller gains, suggesting that awareness of word structure contributes to more robust and semantically coherent language representations. Our approach achieves these benefits without modifying tokenization algorithms or model architectures. By pre-processing the tokenizer's training data with a morphological segmentation model, we guide it towards linguistically sound subword units. This method is easily extendable to other languages, provided a morphological segmenter is available, or in its abscence, a language-agnostic option.

Our results reinforce prior findings that incorporating morphology into tokenization improves model performance and suggest embracing linguistic structure as a way to enrich language modeling. To enable further research on morphological segmentation for Spanish, we release (under Apache-2.0 license) the segmentation resource, reference data, lexicon, and affix list (cf., Section 3), along with the tokenizers and models presented in Sections 3.2 and 5.1, respectively [11].

[11] https://github.com/Albalbalba/morphtokenizer

Limitations

Despite its simplicity and effectiveness, our approach comes with several limitations.

First, the resulting tokenization is not fully morphological. We intentionally avoid highly granular segmentations based on the intuition that they may hinder language modeling. As a result, some frequent derivational and inflectional sequences are grouped together, even if they consist of multiple morphs. Additionally, we do not alter the tokenization algorithm itself, but rather influence it indirectly by training it over segmented data. This makes integration straightforward but limits control: the algorithm sometimes joins morphs that the segmenter would have seprated (like verb inflections and clitics). Concurrent work (Asgari et al., 2025) addresses this by injecting morphological information directly into the algorithm, though we reserve judgment until implementation details become available.

Second, our work focuses on a single language, Spanish, which is morphologically richer than English, but arguably not at the most complex end of the typological spectrum. We expect our method to yield stronger benefits in agglutinative or polysynthetic languages. Moreover, this approach is language-specific, whereas modern LLMs are multilingual. Adapting morphology-aware tokenization for such models is a challenging and open problem that is already attracting attention (see the multilingual approach based on morphological segmentation proposed by Limisiewicz et al. (2024), for instance).

Third, we only experimented with a single masked language model architecture (RoBERTa) and relatively small model sizes by current standards. It is possible that the observed improvements may diminish at larger scales, where models can have the capacity to bypass suboptimal tokenizations. However, even very large models have shown a poor grasp of the underlying morphology (Ismayilzada et al., 2025). Our evaluation is also limited in scope, focusing mostly on sentence-level tasks. While results are promising, morphology-aware tokenization may show clearer benefits for token-level and morphologically sensitive tasks such as POS tagging, parsing or semantic role labeling. Moreover, the high computational cost of pretraining limits our ability to explore larger model sizes as well as the eventual impact of parameters like vocabulary size or masking strategy.

Our focus was not to compete with other pretrained models, however, their performance can provide a useful reference point. Thus, for context, we compared our models with BETO (Canete et al., 2020) and mBERT (Devlin et al., 2019). While both of them outperform our morphologically informed model on downstream tasks (achieving, for instance, an F1 score for PAWS-X of 0.89 vs. 0.85), the opposite is true for tasks that require a better morphological knowledge (the morph-aware model achieves a 0.93 accuracy in the agreement tests compared to BETO's 0.91 or mBERT's 0.80), see Table 9 in the Appendix for further details.

Finally, the fact that word-level tokenization is commonly assumed as a fixed component of the language modeling pipeline, does not mean it is the only option. Recent proposals challenge this assumption altogether suggesting alternatives both at the lower level (Clark et al., 2022) and at the higher one (LCM team, 2024). It would be valuable to investigate how these alternatives fare compared to our suggested approach, particularly in tasks where morphological structure plays a central role.

Acknowledgments

The work presented in this paper has been partially supported by the European Commission in the framework of the Horizon Europe program (contract number 101070278). We sincerely appreciate the anonymous reviewers for their valuable suggestions and thoughtful feedback, which greatly contributed to enhancing the quality of this paper. We also acknowledge the use of the MareNostrum 5 supercomputer at the Barcelona Supercomputing Center (BSC) for model training.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval*

2014), pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Catherine Arnett, Pamela D. Rivière, Tyler A. Chang, and Sean Trott. 2024. Different tokenization schemes lead to comparable performance in Spanish number agreement. In *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–38, Mexico City, Mexico. Association for Computational Linguistics.

Ehsaneddin Asgari, Yassine El Kheir, and Mohammad Ali Sadraei Javaheri. 2025. Morphbpe: A morpho-aware tokenizer bridging linguistic complexity for efficient llm training across morphologies. *Preprint*, arXiv:2502.00894.

Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 55–60, New Orleans. Association for Computational Linguistics.

Thomas Bauwens and Pieter Delobelle. 2024. BPE-knockout: Pruning pre-existing BPE tokenisers with backwards-compatible morphological semi-supervision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5810–5832, Mexico City, Mexico. 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.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. *PML4DC at ICLR*, 2020.

Kenneth Ward Church. 2020. Emerging trends: Subwords, seriously? *Natural Language Engineering*, 26(3):375–382.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith Klavans, and Smaranda Muresan. 2020. MorphAGram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7112–7122, Marseille, France. European Language Resources Association.

Ramy Eskander, Cass Lowry, Sujay Khandagale, Francesca Callejas, Judith Klavans, Maria Polinsky, and Smaranda Muresan. 2021. Minimally-supervised morphological segmentation using Adaptor Grammars with linguistic priors. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3969–3974, Online. Association for Computational Linguistics.

Ramy Eskander, Cass Lowry, Sujay Khandagale, Judith Klavans, Maria Polinsky, and Smaranda Muresan. 2022. Unsupervised stem-based cross-lingual part-of-speech tagging for morphologically rich low-resource languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4061–4072, Seattle, United States. Association for Computational Linguistics.

Ramy Eskander, Owen Rambow, and Smaranda Muresan. 2018. Automatically tailoring unsupervised morphological segmentation to the language. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 78–83, Brussels, Belgium. Association for Computational Linguistics.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016. Extending the use of Adaptor Grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 900–910, Osaka, Japan. The COLING 2016 Organizing Committee.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland. Dublin City University and 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.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.

Haris Jabbar. 2024. Morphpiece : A linguistic tokenizer for large language models. *Preprint*, arXiv:2307.07262.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

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.

Venelin Kovatchev and Mariona Taulé. 2022. InferES : A natural language inference corpus for Spanish featuring negation-based contrastive and adversarial examples. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3873–3884, Gyeongju, Republic of Korea. International Committee on 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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Paul-Ambroise Duquenne Maha Elbayad-Artyom Kozhevnikov Belen Alastruey Pierre Andrews Mariano Coria Guillaume Couairon Marta R. Costa-jussà David Dale Hady Elsahar Kevin Heffernan João Maria Janeiro Tuan Tran Christophe Ropers Eduardo Sánchez Robin San Roman Alexandre Mourachko Safiyyah Saleem Holger Schwenk LCM team, Loïc Barrault. 2024. Large Concept Models: Language modeling in a sentence representation space.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Shu Okabe and François Yvon. 2023. Joint word and morpheme segmentation with Bayesian nonparametric models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 640–654, Dubrovnik, Croatia. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1525–1534, Berlin, Germany. 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.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China. Association for Computational Linguistics.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.

Real Academia Española. 2010. *Nueva gramática de la lengua española: manual*. Asociacion de Academias de la Lengua Española.

Jonne Saleva and Constantine Lignos. 2021. The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Preprint*, arXiv:1508.07909.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 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*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online. Association for Computational Linguistics.

Ngoc Tan Le, Antoine Cadotte, Mathieu Boivin, Fatiha Sadat, and Jimena Terraza. 2022. Deep learning-based morphological segmentation for indigenous languages: A study case on innu-aimun. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 146–151, Hybrid. Association for Computational Linguistics.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Thinh Truong, Yulia Otmakhova, Karin Verspoor, Trevor Cohn, and Timothy Baldwin. 2024. Revisiting subword tokenization: A case study on affixal negation in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5082–5095, Mexico City, Mexico. Association for Computational Linguistics.

Peter ; Grönroos Stig-Arne ; Kurimo Mikko Virpioja, Sami ; Smit. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. In *Aalto University publication series*. Department of Signal Processing and Acoustics, Aalto University.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

Matheus Westhelle, Luciana Bencke, and Viviane P. Moreira. 2022. Impact of morphological segmentation on pre-trained language models. In *Intelligent Systems*, pages 402–416, Cham. Springer International Publishing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and 12 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yi Zhu, Ivan Vulić, and Anna Korhonen. 2019. A systematic study of leveraging subword information for learning word representations. 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 912–932, Minneapolis, Minnesota. Association for Computational Linguistics.

A Additional Comparisons

Our objective was not to outperform existing pretrained models, yet their results provide a useful benchmark. To offer perspective, we compared our models with BETO (Canete et al., 2020) and mBERT (Devlin et al., 2019), c.f. Table 9. While these models achieve higher scores on standard downstream tasks, our morphologically informed model shows superior performance on agreement tests, which specifically require stronger morphological knowledge.

Model	**XNLI** (Acc.)	**PAWS-X** (F1)	**STS** (Comb.)	**Agreement** (Acc.)
Morph-aware	0.742	0.845	0.801	**0.926**
Vanilla	0.733	0.841	0.776	0.864
BETO	**0.813**	**0.893**	**0.816**	0.913
mBERT	0.771	0.886	0.807	0.795

Table 9: Comparison to existing baselines