Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Using Morphological Lexicons for Weakly Supervised Subword Vocabulary Learning and Bias-Aware Evaluation

Yifei Li
Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University
Zijun Li
College of Electronic and Information Engineering, Tongji University
Hui Xiao
College of Electronic and Information Engineering, Tongji University
Xuefeng Li
lixuefeng@tongji.edu.cn
College of Electronic and Information Engineering, Tongji University

Research Article

Keywords: Morphological lexicon, Subword tokenization, Weak supervision, Vocabulary learning, Morpheme-boundary evaluation, Bias-aware evaluation

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10260345/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Using Morphological Lexicons for Weakly Supervised Subword Vocabulary Learning and Bias-Aware Evaluation

Yifei Li[1], Zijun Li[2], Hui Xiao[1,2], Xuefeng Li[1,2*]

[1]Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China.
[2]College of Electronic and Information Engineering, Tongji University, Shanghai, China.

*Corresponding author(s). E-mail(s): lixuefeng@tongji.edu.cn;
Contributing authors: 2565917074lyf@gmail.com; lizijun@tongji.edu.cn; xiaohui@mail.tongji.edu.cn;

Abstract

Subword vocabularies are usually learned from corpus statistics. This is useful for coverage, but it can produce units that cross morpheme boundaries and make the resulting vocabulary less transparent. This article studies a simple use of a morphological lexicon in subword vocabulary learning. We convert MorphoLex into prefix, root/stem, and suffix inventories, and use these inventories as weak supervision during the training of BPE, WordPiece-style, and Unigram tokenizers. The change is limited to vocabulary learning: the learned tokenizer is applied with the standard inference procedure of the base method. The same resource also motivates a bias-aware evaluation protocol, with controls for affix-completeness bias and for word-level provenance overlap between the lexicon used for guidance and the words used for evaluation. On MorphoLex-derived English test sets, lexicon-guided training improves morpheme-boundary F1 by +5.00 pp for BPE, +3.23 pp for Unigram, and +1.10 pp for WordPiece-style training at the best tested settings. The gains persist when the affix-completeness constraint is removed and when guidance and evaluation use disjoint MorphoLex word pools. The results support a modest but useful claim: morphological lexicons can serve both as weak supervision for tokenizer training and as a resource for controlled evaluation.

Keywords: Morphological lexicon, Subword tokenization, Weak supervision, Vocabulary learning, Morpheme-boundary evaluation, Bias-aware evaluation

1 Introduction

Subword tokenization is now a routine part of NLP systems. Byte Pair Encoding (BPE) [1–3], WordPiece [4], and the Unigram language model [5] give practical open-vocabulary coverage while keeping sequence lengths manageable. Their vocabularies, however, are learned mainly from corpus-level statistics. Frequent or probable character sequences may be selected even when they cut across morpheme boundaries. For morphologically complex words, this can hide relations among prefixes, roots, and suffixes and reduce the linguistic interpretability of the vocabulary.

Many language resources already contain the kind of information that statistical tokenizers lack. Morphological lexicons such as MorphoLex [6], CELEX [7], and UniMorph [8] record roots, affixes, inflectional patterns, or derivational structure. These resources are often used for analysis or evaluation, but they are less often used as lightweight supervision for mainstream tokenizer training. A hard use of such resources would require full boundary annotation for the training corpus. That requirement is unattractive for large raw-text corpora. A softer use is to turn a lexicon into a prior: candidate subwords that match known morphemes are preferred, but the base tokenizer remains driven by corpus evidence.

This article takes the latter approach. We propose Lex+X, a lexicon-guided variant of subword vocabulary learning. A morphological lexicon is converted into three inventories: prefixes, roots or stems, and suffixes. During vocabulary learning, a candidate unit receives a small multiplicative bonus if it matches one of these inventories under the relevant boundary condition. The same idea is applied to BPE, WordPiece-style training, and Unigram. The method changes only the vocabulary learning stage. It does not require gold morpheme boundaries for the tokenizer training corpus, and it does not change inference-time decoding.

The second part of the article is evaluative. If the same lexicon supplies both the weak supervision signal and the gold morphology used for testing, improvements may reflect resource overlap rather than better morphological segmentation. This is a familiar problem in language resource evaluation: the resource is useful, but its role must be made explicit. We therefore construct two diagnostics. The first removes the affix-completeness constraint from the main evaluation set. The second separates MorphoLex word entries into a source pool for inventory construction and a held-out pool for evaluation. These tests do not remove all morpheme overlap, which would ask a different question, but they do control two concrete sources of evaluation bias.

The contribution is therefore not a new tokenizer family. It is a resource-guided learning and evaluation setup for existing tokenizers. The study asks two questions: can a morphological lexicon improve boundary alignment in standard subword vocabulary learning, and can such improvement be evaluated without relying only on a single resource-overlapping test set? Experiments on English word forms from WikiText-103 and MorphoLex show positive gains for all three tokenizer families. BPE benefits most, Unigram benefits substantially, and WordPiece-style training shows smaller but consistent gains. The pattern suggests that the same morphological resource interacts differently with frequency-driven, association-driven, and probability-driven learners.

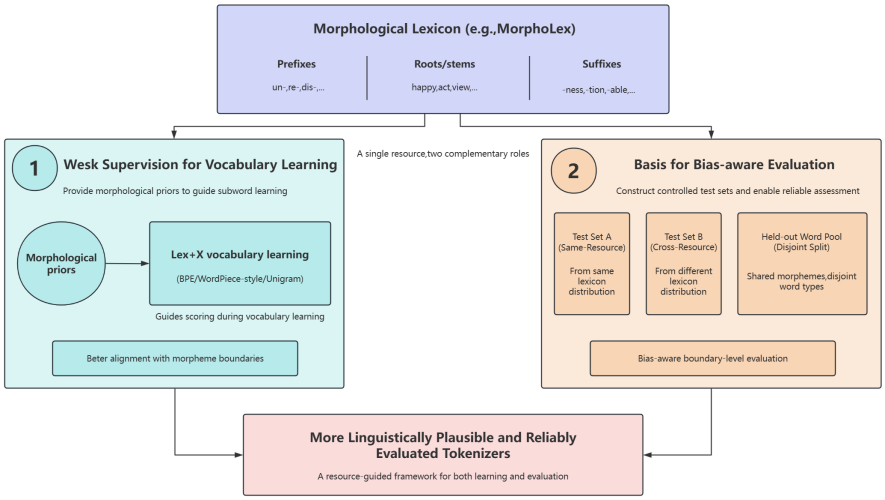


Fig. 1 Dual role of the morphological lexicon. The same resource provides weak supervision for vocabulary learning and supports bias-aware evaluation, but the two uses are separated in the experimental protocol.

Figure 1 summarizes the resource role in the study. MorphoLex is used both as weak supervision for vocabulary learning and as the basis for controlled evaluation. This dual role is central to the paper and to its fit with language resources and evaluation.

2 Related Work and Positioning

2.1 Subword Tokenization and Vocabulary Learning

Subword tokenization provides a practical solution to the open-vocabulary problem in NLP. Byte Pair Encoding (BPE) was first introduced as a compression algorithm [1], later adapted to speech and text processing [2], and widely adopted in neural machine translation after [3]. BPE constructs a vocabulary through greedy adjacent-symbol merges, and its simplicity has made it a common choice in large-scale pretrained models such as GPT-2 [9] and RoBERTa [10]. WordPiece [4], used in BERT [11], follows a related merge-based strategy but ranks candidate units by an association-oriented criterion rather than raw frequency. The Unigram language model [5] takes a probabilistic view, estimating a distribution over candidate subwords and pruning the vocabulary through likelihood-based optimization.

These approaches share a common reliance on corpus-internal statistical signals. Their vocabularies are learned from frequency, association, or likelihood patterns rather than from explicit linguistic structure. Several extensions have been proposed to address this limitation. Subword regularization [12] samples alternative segmentations to improve robustness. BPE-Dropout [13] introduces stochasticity into merge application. Studies on vocabulary size [14] show that tokenization granularity can substantially affect downstream performance. Tokenization-free or byte-level approaches such as ByT5 [15] and CANINE [16] reduce dependence on learned vocabularies but change

the modeling regime. A growing body of work therefore treats tokenization not as a neutral preprocessing step, but as a representational decision that shapes the units available to the model [17–19]. The present study shares this perspective: the goal is not to replace mainstream tokenizers, but to examine whether their vocabulary learning objectives can be guided by morphological knowledge.

2.2 Morphological Resources and Morphologically Informed Tokenization

Morphological structure has been studied through both lexical resources and computational models. MorphoLex [6] provides derivational analyses for approximately 70,000 English words. CELEX [7] offers detailed lexical and morphological information for English and other languages. UniMorph [8] provides a multilingual schema for inflectional morphology. These resources encode structure at the lexical level and have supported shared tasks in morphological analysis such as Morpho Challenge and SIGMORPHON evaluations, which emphasize standardized benchmarks and transparent evaluation protocols.

Several lines of work have explored how morphological information can improve tokenization or subword modeling. Constraint-based methods such as MorphoBPE restrict merges that cross known morpheme boundaries, but they require boundary-annotated training data, which limits scalability. Compositional morphology has been used to improve word representations and language modeling [20]. Studies have examined how neural models capture morphological information across character-to-word boundaries [21] and how morphologically informed representations benefit machine translation [22]. Derivational morphology has been shown to affect how pretrained language models interpret complex words [23]. Most recently, LEGROS [24] integrates lexical semantic information into subword segmentation through Morfessor-based pretokenization and a subword bigram model. Its design, however, introduces additional modeling components. Lex+X differs by using a lightweight lexicon lookup as a multiplicative bonus within the standard training objective of each base tokenizer, without requiring pretrained word vectors, a separate segmentation model, or a new inference procedure. The broader distinction is between morphology as a hard segmentation constraint and morphology as a weak vocabulary-learning prior. Hard constraints require reliable boundary information for the training data; weak supervision allows a lexical resource to guide learning without fully determining the segmentation.

2.3 Evaluation of Morphology-Aware Tokenization

Evaluation is a central issue in morphology-aware tokenization. Boundary-level precision, recall, and F1 provide a direct measure of whether predicted subword boundaries align with gold morpheme boundaries. These metrics are useful for intrinsic evaluation, but they are not sufficient on their own. Recent work shows that segmentation consistency across inflected forms and word classes can be as important as aggregate boundary adherence [25]. Tokenization choices can also influence downstream behavior and model efficiency in ways not fully captured by boundary-level metrics.

For lexicon-guided methods, evaluation raises an additional resource-specific concern. If the same lexical resource supplies both the guidance signal and the gold evaluation structure, performance gains may partly reflect overlap between the resource used for supervision and the test set. A test set that deliberately covers all affixes in the guiding lexicon may inflate the apparent effect of the method, and a test set drawn from the same database may share morphological distributions with the resource used for supervision even when word forms do not overlap. These issues do not invalidate lexicon-based evaluation, but they require explicit diagnostics. The present study therefore treats evaluation design as part of the contribution. In addition to standard boundary metrics, the evaluation includes analyses by vocabulary size, morpheme type, and word complexity, along with two robustness diagnostics: a completeness-bias analysis that removes the affix-coverage constraint, and a word-level provenance analysis that separates the entries used for guidance from those used for evaluation. This design aligns morphology-aware tokenization with the broader concerns of language resource evaluation: transparent resource use, controlled test construction, and interpretable system comparison.

3 Lex+X: Lexicon-Guided Vocabulary Learning

Lex+X is a training-time modification of existing subword vocabulary learners. It does not replace the base tokenizer. It adds a morpheme-based preference to candidate scoring and then uses the learned vocabulary with the standard inference procedure of the base method.

3.1 Resource Representation

Let the word-level training corpus be $\mathcal{D} = \{w_1, \ldots, w_N\}$. A subword vocabulary learner constructs a vocabulary $\mathcal{V}$ of target size K so that each word can be segmented as

$$w = t_1 t_2 \cdots t_m, \qquad t_i \in \mathcal{V}. \tag{1}$$

Lex+X adds a morphological resource

$$\mathcal{L} = (\mathcal{P}, \mathcal{R}, \mathcal{S}), \tag{2}$$

where $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{S}$ are prefix, root/stem, and suffix inventories extracted from a lexicon. In this study the lexicon is MorphoLex. After normalization and filtering, the inventory contains 105 prefixes, 4,999 roots/stems, and 106 suffixes.

The lexicon is not used as a complete parser. A candidate unit is simply checked against the relevant inventory. Missing entries do not block training, and unmatched candidates remain available to the base tokenizer.

3.2 Boundary-Aware Matching

Affix status is position-sensitive. To make this visible during vocabulary learning, each word is temporarily marked with boundary symbols:

$$w \to \hat{}\, w\, \$. \tag{3}$$

A candidate adjacent to the left boundary can match the prefix inventory, a candidate adjacent to the right boundary can match the suffix inventory, and a candidate inside the word can match the root/stem inventory. Boundary symbols are removed from final output tokens. They are used only to score candidates during vocabulary learning (Eq. 3).

The binary matching function is

$$\mathrm{Morph}(x) = \begin{cases} 1, & \text{lexicon match under boundary condition,} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This intentionally simple design keeps the resource interface transparent. It does not model allomorphy, morpheme ambiguity, or full word-level analyses.

3.3 Morphological Bonus

For any candidate merge or subword candidate c, the morphology-enhanced score is

$$\mathrm{Score}_{\mathrm{Lex+X}}(c) = \mathrm{Score}_{\mathrm{base}}(c)\,(1 + \lambda\,\mathrm{Morph}(c))\,, \tag{5}$$

where $\lambda \geq 0$ controls the strength of the lexicon-derived preference and $\mathrm{Morph}(c)$ is defined in Eq. 4. At $\lambda = 0$, the standard tokenizer is recovered exactly.

For BPE, $c = a \oplus b$ is the unit created by merging adjacent symbols a and b:

$$\mathrm{Score}_{\mathrm{LexBPE}}(a, b) = f(a, b)\,(1 + \lambda_{\mathrm{BPE}}\mathrm{Morph}(a \oplus b))\,. \tag{6}$$

For WordPiece-style training we use a reproducible association score,

$$\mathrm{Score}_{\mathrm{LexWP}}(a, b) = \frac{f(a, b)}{f(a)f(b)}\,(1 + \lambda_{\mathrm{WP}}\mathrm{Morph}(a \oplus b))\,. \tag{7}$$

This should be read as a transparent WordPiece-style implementation, not as a reconstruction of the original proprietary WordPiece trainer. For Unigram, the bonus is applied after the usual probability update:

$$p'(x) \propto p(x)\,(1 + \lambda_{\mathrm{Uni}}\mathrm{Morph}(x))\,, \qquad \sum_{x \in \mathcal{V}} p'(x) = 1. \tag{8}$$

3.4 Training and Inference

Figure 2 gives the full pipeline. Boundary-aware preprocessing is applied before training. The morpheme inventories affect candidate scoring. The final vocabulary is then used

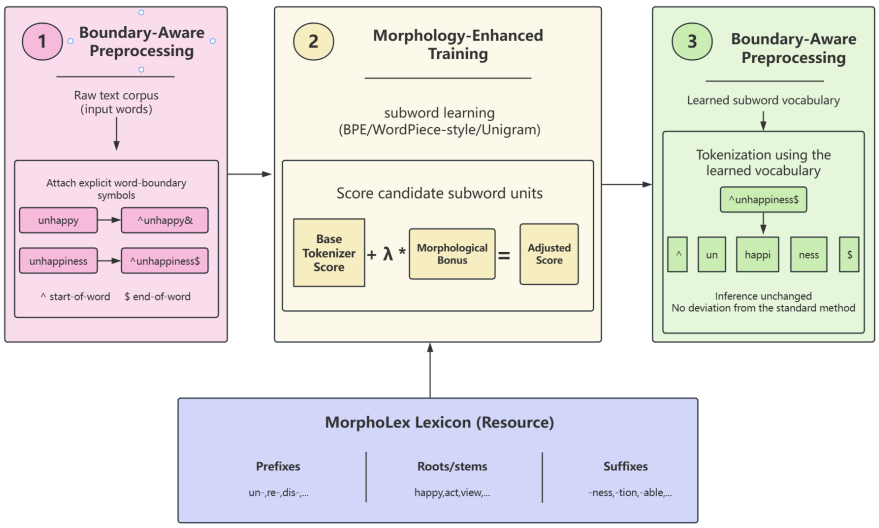


Fig. 2 Lex+X pipeline. Boundary symbols make word edges visible during vocabulary learning. MorphoLex-derived inventories affect only candidate scoring. Standard inference is unchanged.

Algorithm 1 Lex+X training for merge-based tokenizers

Require: Word corpus $\mathcal{D}$, target size K, inventories $\mathcal{L}$, base score $\text{Score}_{\text{base}}$, weight λ
Ensure: Vocabulary $\mathcal{V}$

Initialize $\mathcal{V}$ with characters and boundary symbols
Convert each word to its boundary-aware representation
while $|\mathcal{V}| < K$ **do**
    Count adjacent candidate pairs (a, b)
    For each candidate, set $c \leftarrow a \oplus b$
    Compute $s(c) = \text{Score}_{\text{base}}(a, b)(1 + \lambda \text{Morph}(c))$
    Select the pair with the highest $s(c)$
    Merge the selected pair and add the merged unit to $\mathcal{V}$
end while
return $\mathcal{V}$

by the base tokenizer in the usual way: merge-rule application for BPE, longest-match decoding for WordPiece-style tokenization, and Viterbi segmentation for Unigram.

Algorithm 1 sketches the merge-based version. For Unigram, the same lookup-and-bonus step is inserted into the probability update and pruning loop. The additional cost is one hash-set lookup per candidate merge or candidate subword. Inference requires no lexicon.

Table 1 Resource and evaluation settings.

Component	Scale	Role
MorphoLex entries	∼70,000 words	Source of morphological analyses and lexicon-derived weak supervision.
Prefix inventory	105 prefixes	Word-initial guidance.
Root/stem inventory	4,999 roots/stems	Lexical guidance independent of affix position.
Suffix inventory	106 suffixes	Word-final guidance.
Test Set A	2,000 words	Main POS-stratified evaluation set with affix-completeness constraint.
Test Set B	2,000 words	Diagnostic set without the affix-completeness constraint.
Held-out split	80% / 20%	Source entries for inventory construction are separated from evaluation entries.
Boundary metrics	P/R/F1	Intrinsic alignment between predicted subword boundaries and gold morpheme boundaries.

4 Resource, Data, and Evaluation Protocol

This section records the resource processing and evaluation design. The details matter because MorphoLex is used both to derive the weak supervision signal and to construct morphology-oriented test sets.

4.1 MorphoLex Processing

MorphoLex provides derivational analyses for approximately 70,000 English word forms [6]. We extract prefixes, roots or stems, and suffixes (Eq. 2) from the segmented analyses. Strings are lowercased; boundary markers and special formatting symbols are removed; duplicates are merged; and very low-frequency affix candidates are filtered to reduce idiosyncratic labels. The resulting inventory has 5,210 morpheme entries in total.

The inventory is used only for candidate matching. It is not used to annotate WikiText-103, and it is not required at inference time. This distinction is important: the method uses MorphoLex as weak supervision, not as an oracle segmentation system.

4.2 Evaluation Sets and Boundary Metrics

The main evaluation set, Test Set A, contains 2,000 unique MorphoLex word forms with gold morphological segmentations. Sampling is part-of-speech stratified. An affix-completeness constraint ensures that every extracted prefix and suffix appears at least once in the evaluation set. The goal is broad affix coverage, not a claim that this is the only natural test distribution.

For a predicted subword sequence $t_1 \dots t_m$ (Eq. 1) and a gold morpheme sequence $g_1 \dots g_n$, predicted and gold boundary sets are

$$B_{\text{pred}} = \left\{ \sum_{i=1}^{j} |t_i| \;\middle|\; j = 1, \dots, m-1 \right\}, \quad B_{\text{gold}} = \left\{ \sum_{i=1}^{j} |g_i| \;\middle|\; j = 1, \dots, n-1 \right\}. \quad (9)$$

Precision, recall, and F1 are computed from $B_{\text{pred}} \cap B_{\text{gold}}$ (Eq. 9) and micro-averaged over words.

Two diagnostic evaluations are added. Test Set B is sampled with the same POS-stratified procedure but without affix completeness. It also contains 2,000 words and shares 89 word forms with Test Set A. For provenance control, MorphoLex word entries are split into an 80% source pool and a 20% held-out pool. The guiding inventory is derived only from the source pool, and the held-out evaluation set is sampled only from the held-out pool. This creates zero word-entry overlap. The split is deterministic (seed 42). It does not create a morpheme-disjoint split: productive morphemes remain shared across pools, with overlap rates of 93% for roots, 96% for prefixes, and 95% for suffixes.

Evaluation by word complexity is also reported. Words with one or two gold morphemes are treated as simple; words with three or more gold morphemes are treated as complex.

4.3 Tokenizer Training Corpus and Settings

Tokenizer vocabularies are learned from WikiText-103. The raw corpus is cleaned by removing empty lines, normalizing whitespace, filtering non-English characters while retaining letters, digits, hyphens, and common punctuation, and discarding sentences shorter than 5 or longer than 100 words. From the cleaned corpus, 500,000 sentences are sampled. Words are lowercased; alphabetic and hyphenated English forms are retained; words shorter than four characters and forms with corpus frequency below 20 are removed. The resulting training corpus contains about 150,000 distinct word forms. MorphoLex gold analyses are not used to annotate this corpus. All random sampling across the experiments uses a fixed seed of 42 for reproducibility.

The systems compared are BPE versus LexBPE, WordPiece-style training versus LexWordPiece, and Unigram versus LexUnigram. Standard baselines are obtained by setting $\lambda = 0$ in the same implementation framework. BPE is evaluated from 9k to 17k vocabulary entries; WordPiece-style and Unigram are evaluated from 5k to 13k. The main guidance weights are $\lambda = 2.5$ for LexBPE, $\lambda = 2.4$ for LexWordPiece, and $\lambda = 2.0$ for LexUnigram. Nearby values are checked as a sensitivity analysis. For Unigram, the initial seed vocabulary is $1.15K$, with four nearby alternatives (± 50, ± 100); the maximum subword length is 10 and the maximum number of EM iterations is 20.

All tokenizers (BPE, WordPiece-style, and Unigram) are implemented from scratch in pure Python without external tokenizer libraries. Morpheme inventories are stored as hash sets. On the reported machine, a single BPE run takes about 8–12 minutes, a WordPiece-style run 4–15 minutes, and a Unigram run about 8 seconds. The Lex+X overhead is small relative to corpus scanning and vocabulary optimization.

Table 2 Main boundary-alignment results at the best tested settings. Each P/R/F1 cell reports precision, recall, and F1.

Family	Vocab.	λ	Standard P/R/F1	Lex+X P/R/F1	ΔF1
BPE	13k	2.5	0.3168 / 0.3588 / 0.3365	0.3653 / 0.4103 / 0.3865	+5.00
WordPiece-style	9k	2.4	0.1651 / 0.8945 / 0.2788	0.1812 / 0.7238 / 0.2898	+1.10
Unigram	9k	2.0	0.3402 / 0.4857 / 0.4001	0.3655 / 0.5292 / 0.4324	+3.23

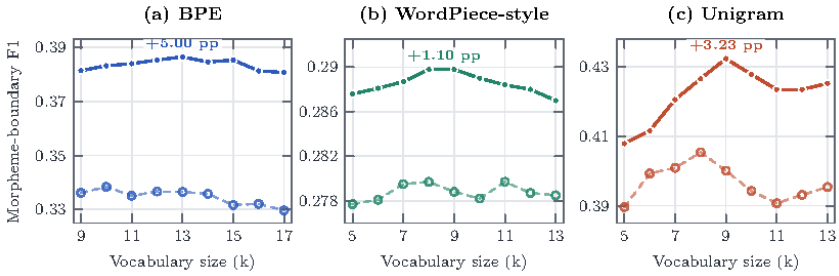


Fig. 3 Morpheme-boundary F1 across vocabulary sizes. Solid lines denote Lex+X variants; dashed lines denote the corresponding standard tokenizers. The y-axis ranges differ across panels to make within-family changes visible.

5 Results

5.1 Boundary Alignment Across Tokenizer Families

Table 2 reports the best settings for the three tokenizer families. Lex+X (Eq. 5) improves morpheme-boundary F1 for each base method. LexBPE (Eq. 6) shows the largest gain, from 0.3365 to 0.3865 at 13k vocabulary entries (+5.00 pp). LexUnigram (Eq. 8) improves from 0.4001 to 0.4324 at 9k (+3.23 pp). LexWordPiece (Eq. 7) improves more modestly, from 0.2788 to 0.2898 at 9k (+1.10 pp).

The precision–recall behavior differs by family. LexBPE and LexUnigram improve both precision and recall at their best settings. LexWordPiece improves precision but lowers recall, giving a small net F1 gain. This is consistent with a more selective segmentation pattern under the association-based WordPiece-style score.

Figure 3 shows the same comparison across vocabulary sizes. The Lex+X curve stays above the corresponding standard curve throughout the tested range for all three families. The peaks are broad rather than isolated, especially for BPE and Unigram. The result is therefore not tied to a single vocabulary size.

Sensitivity to λ is limited. For LexBPE at 13k, F1 rises from the $\lambda = 0$ baseline and remains near its peak for $\lambda = 2.5$–2.9. For LexWordPiece at 9k, $\lambda = 2.1$–2.7 gives F1 values between 0.2883 and 0.2898. For LexUnigram at 9k, $\lambda = 1.6$–2.4 gives F1 values between 0.4295 and 0.4322. The parameter therefore behaves as a coarse balance between corpus statistics and morphological preference, not as a brittle tuning constant.

Table 3 Compact summary of λ sensitivity at the main vocabulary size for each family.

Family	Vocab.	Best λ	Baseline F1	Observed stability
BPE	13k	2.5	0.3365	F1 remains 0.3848–0.3865 for $\lambda = 2.5$–2.9.
WordPiece-style	9k	2.4/2.5	0.2788	F1 remains 0.2883–0.2898 for $\lambda = 2.1$–2.7.
Unigram	9k	2.0	0.4001	F1 remains 0.4295–0.4322 for $\lambda = 1.6$–2.4.

Table 4 Contribution of morpheme types. Values are F1 scores at the main vocabulary size for each tokenizer family.

Family	Guidance	# morphemes	F1	ΔF1
BPE	none	–	0.3365	–
BPE	prefixes only	105	0.3408	+0.43
BPE	roots/stems only	4,999	0.3711	+3.46
BPE	suffixes only	106	0.3460	+0.95
BPE	all	5,210	0.3865	+5.00
WordPiece-style	none	–	0.2788	–
WordPiece-style	prefixes only	105	0.2819	+0.31
WordPiece-style	roots/stems only	4,999	0.2785	-0.03
WordPiece-style	suffixes only	106	0.2852	+0.64
WordPiece-style	all	5,210	0.2898	+1.10
Unigram	none	–	0.4001	–
Unigram	prefixes only	105	0.3964	-0.37
Unigram	roots/stems only	4,999	0.4202	+2.01
Unigram	suffixes only	106	0.3996	-0.05
Unigram	all	5,210	0.4324	+3.23

5.2 Morpheme-Type Contributions

Table 4 separates the effect of prefixes, roots/stems, and suffixes. BPE and Unigram gain most from roots/stems. WordPiece-style training gains most from suffixes, while root-only guidance adds little. The full inventory is best for all three families, even when individual morpheme types are weak or slightly negative in isolation.

The result argues against treating inventory size as the only measure of resource value. Roots/stems dominate for two tokenizers, but the smaller affix inventories still matter in combination. The usefulness of a morpheme class depends on the base learner's scoring bias.

Table 5 F1 by word complexity. Simple words have one or two gold morphemes; complex words have three or more.

Family	Simple			Complex		
	Std.	Lex+X	Δ	Std.	Lex+X	Δ
BPE	0.3538	0.4038	+5.00	0.3131	0.3631	+5.00
WordPiece-style	0.2354	0.2543	+1.89	0.3470	0.3473	+0.03
Unigram	0.3848	0.4133	+2.85	0.4205	0.4585	+3.80

Table 6 Bias-aware diagnostics. Bias is $\Delta F1_A - \Delta F1_B$. Retention is held-out ΔF1 divided by original ΔF1.

Family	A Δ	B Δ	Bias	Held-out Δ	Ret.
BPE	+5.00	+4.95	+0.05	+5.29	105.8%
WordPiece-style	+1.10	+1.17	-0.07	+0.73	66.4%
Unigram	+3.23	+3.14	+0.09	+3.64	112.7%

5.3 Word Complexity

Table 5 compares simple words (one or two gold morphemes; 1,364 samples) with complex words (three or more gold morphemes; 636 samples). LexBPE gains +5.00 pp in both groups. LexWordPiece gains mainly on simple words. LexUnigram gains in both groups and reaches the highest complex-word F1, 0.4585.

The WordPiece-style pattern is informative. Its standard baseline is already much stronger on complex words than on simple words, leaving little room for the lexicon bonus in that subgroup. BPE has the opposite profile: a simple frequency objective leaves substantial room for resource guidance. Unigram combines a strong baseline with a clear lexicon gain, especially on complex words.

5.4 Bias-Aware Diagnostics

Table 6 gives the two evaluation-bias checks. Removing the affix-completeness constraint changes the measured gain by less than 0.1 pp for all families. The sign is not consistent across families, so the constraint does not appear to systematically inflate Lex+X performance.

The held-out word-pool test also preserves positive gains. BPE and Unigram retain slightly more than the original gain; WordPiece-style training retains about two thirds of its smaller original gain. This held-out protocol controls word-entry provenance, not morpheme identity. The result should therefore be interpreted narrowly: the method does not depend on reusing the same MorphoLex word entries for guidance and evaluation.

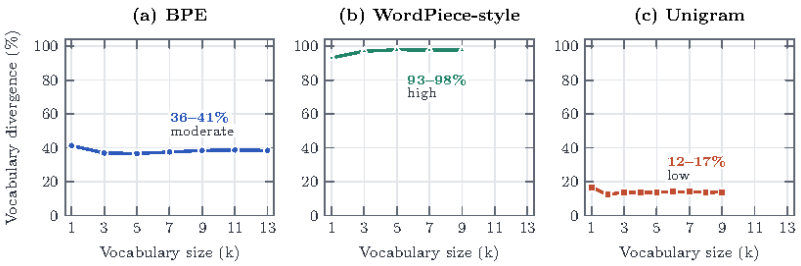


Fig. 4 Vocabulary divergence between each standard tokenizer and its Lex+X counterpart.

Table 7 Summary of vocabulary divergence ranges across the tested vocabulary sizes.

Family	Divergence range	Best ΔF1	Reading
BPE	36.6–41.3%	+5.00	Moderate vocabulary change with the largest boundary gain.
WordPiece-style	93.1–97.9%	+1.10	Large vocabulary change, but limited boundary improvement.
Unigram	12.4–16.6%	+3.23	Small vocabulary change with a clear boundary gain.

5.5 Vocabulary Divergence

A final diagnostic asks how much the learned vocabulary changes. For a standard vocabulary $\mathcal{T}_{\text{std}}$ and a Lex+X vocabulary $\mathcal{T}_{\text{lex}}$, divergence is defined as one minus Jaccard similarity:

$$D(\mathcal{T}_{\text{std}}, \mathcal{T}_{\text{lex}}) = 1 - \frac{|\mathcal{T}_{\text{std}} \cap \mathcal{T}_{\text{lex}}|}{|\mathcal{T}_{\text{std}} \cup \mathcal{T}_{\text{lex}}|}. \tag{10}$$

Figure 4 and Table 7 report the divergence values (Eq. 10) and show that vocabulary change and boundary gain are not the same thing. WordPiece-style training has very high divergence, about 93–98%, but the smallest F1 gain. Unigram has low divergence, about 12–17%, while retaining a substantial gain. BPE sits between them, with moderate divergence and the largest boundary improvement.

The diagnostic matters because a resource-guided tokenizer can change many tokens without becoming more morphologically aligned. Boundary-level evaluation remains necessary. Divergence is useful mainly for explaining how each tokenizer family internalizes the same lexicon-derived signal.

6 Discussion

The results support a limited and resource-centered conclusion. A morphological lexicon can guide subword vocabulary learning without serving as a full analyzer and without annotating the training corpus. This is a practical role for resources such as MorphoLex: they can provide a compact inventory that nudges vocabulary learning toward linguistically meaningful units while leaving corpus statistics in control.

The three tokenizer families respond differently. BPE has the weakest built-in structural preference and benefits most from the external prior. WordPiece-style training already uses an association score that captures some internal cohesion, so the added lexicon signal produces a smaller F1 gain and a sharper precision–recall trade-off. Unigram absorbs the prior through probability reweighting, yielding a strong gain with little vocabulary divergence. These differences matter for evaluation. There is no single "morphology-aware tokenization" effect independent of the base learner.

The evaluation protocol is as important as the scoring change. Because MorphoLex supplies both the guidance inventory and the gold analyses, a single resource-overlapping test set would be weak evidence. The completeness-bias and provenance diagnostics make the claim more specific. The gains are not explained by the affix-completeness constraint, and they persist when guidance and evaluation use disjoint word-entry pools. At the same time, productive morphemes remain shared across pools. The held-out test should not be read as evidence of generalization to unseen morphemes.

Several limitations remain. The study uses English derivational morphology and one primary lexicon. Languages with richer inflection, templatic morphology, compounding, reduplication, or non-concatenative processes may need a different matching function. The current binary lookup also ignores allomorphy and morpheme ambiguity. A string that is a suffix in one context and part of a root in another is handled only by simple positional rules. Cross-resource validation with CELEX, UniMorph, DeriNet, MorphyNet, or comparable lexicons would test generality more directly.

The evaluation is intrinsic. Better morpheme-boundary alignment does not by itself prove improved language modeling, translation, classification, or LLM performance. No such downstream claim is made here. The WordPiece-style results should also be read cautiously: the implementation uses a transparent association-score formulation, not the original proprietary WordPiece trainer. Finally, the present tables do not include bootstrap confidence intervals. The repeated vocabulary-size checks, λ sensitivity, completeness-bias analysis, held-out word-pool analysis, and divergence diagnostic all point in the same direction, but paired resampling would be useful, especially for the small WordPiece-style gain.

For language-resource evaluation, the broader lesson is simple. When a resource is used to guide learning, the evaluation should make resource overlap visible. Reporting only a best F1 score hides too much. Reporting boundary alignment, test-set construction controls, provenance controls, and vocabulary divergence gives a more usable account of what the resource contributes.

7 Conclusion

This article studied a lightweight use of morphological lexicons in subword vocabulary learning. MorphoLex was converted into prefix, root/stem, and suffix inventories, and these inventories were used as weak supervision for BPE, WordPiece-style, and Unigram training. The modification affects only vocabulary learning; standard inference is unchanged. On MorphoLex-derived evaluation sets, Lex+X improves morpheme-boundary F1 for all three tokenizer families, with the largest gain for BPE, a substantial gain for Unigram, and a smaller gain for WordPiece-style training.

The evaluation protocol is part of the contribution. Completeness-bias and word-level provenance diagnostics show that the main gains are not artifacts of a single affix-complete, resource-overlapping test set. Vocabulary divergence further shows that different tokenizer families can change their vocabularies in very different ways under the same resource signal. Morphological lexicons can therefore be used not only as static linguistic resources, but also as controlled weak supervision for tokenizer training and as the basis for more reliable tokenizer evaluation.

Declarations

Funding This work was supported by the National Natural Science Foundation of China under Grant No. 62175187.

Conflicts of interest The authors have no conflicts of interest to declare.

Availability of data and materials WikiText-103 is publicly available. MorphoLex is used under its original distribution terms. The released materials provide scripts and split identifiers for reconstructing the morpheme inventories and evaluation sets from the original MorphoLex distribution.

Code availability The code repository is available at https://github.com/lyf020626/Using-Morphological-Lexicons. The release includes scripts for corpus preprocessing, morpheme-inventory extraction, tokenizer training, boundary-level evaluation, vocabulary-divergence analysis, configuration files, and random seeds.

Author contributions Li Yifei wrote the main manuscript text, implemented the method, and performed the experiments, validation, and analysis. Li Zijun contributed to the research idea, methodological design, and technical implementation. Li Xuefeng conceived the research topic, led the project, guided the methodological direction, and contributed to the discussion and interpretation of the results. Xiao Hui provided comments on the research presentation and manuscript. All authors reviewed and approved the final manuscript.

References

[1] Gage, P.: A new algorithm for data compression. The C Users Journal **12**(2), 23–38 (1994)

[2] Schuster, M., Nakajima, K.: Japanese and korean voice search. In: Proceedings of ICASSP (2012)

[3] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of ACL (2016)

[4] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system:

Bridging the gap between human and machine translation. arXiv:1609.08144 (2016)

[5] Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of EMNLP (2018)

[6] Sánchez-Gutiérrez, R., Mailhot, H., Deacon, S.H., Wilson, M.A.: MorphoLex: A derivational morphological database for 70,000 English words. Behavior Research Methods **50**, 1568–1580 (2018)

[7] Baayen, R.H., Piepenbrock, R., Gulikers, L.: The CELEX Lexical Database (Release 2). Linguistic Data Consortium, Philadelphia (1995)

[8] Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A.D., Kann, K., Mielke, S.J., Nicolai, G., Silfverberg, M., Yarowsky, D., Eisner, J., Hulden, M.: UniMorph 2.0: Universal morphology. In: Proceedings of LREC (2018)

[9] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. Technical report, OpenAI (2019)

[10] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 (2019)

[11] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT (2019)

[12] Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. In: Proceedings of ACL (2018)

[13] Provilkov, I., Emelianenko, D., Voita, E.: BPE-dropout: Simple and effective subword regularization. In: Proceedings of ACL (2020)

[14] Gowda, T., May, J.: Finding the optimal vocabulary size for neural machine translation. In: Findings of EMNLP (2020)

[15] Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., Raffel, C.: ByT5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics **10**, 291–306 (2022)

[16] Clark, J.H., Garrette, D., Turc, I., Wieting, J.: CANINE: Pre-training an efficient tokenization-free encoder for language representation. Transactions of the Association for Computational Linguistics **10**, 73–91 (2022)

[17] Mielke, S.J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si,

C., Lee, W.Y., Sagot, B., Tan, S.: Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. arXiv:2112.10508 (2021)

[18] Bostrom, K., Durrett, G.: Byte pair encoding is suboptimal for language model pretraining. In: Findings of EMNLP (2020)

[19] Schmidt, C.W., Reddy, V., Zhang, H., Alameddine, A., Uzan, O., Pinter, Y., Tanner, C.: Tokenization is more than compression. In: Proceedings of EMNLP, pp. 678–702 (2024)

[20] Botha, J., Blunsom, P.: Compositional morphology for word representations and language modelling. In: Proceedings of ICML (2014)

[21] Vania, C., Lopez, A.: From characters to words to in between: Do we capture morphology? In: Proceedings of ACL (2017)

[22] Ataman, D., Federico, M.: Compositional representation of morphologically-rich input for neural machine translation. In: Proceedings of ACL (2018)

[23] Hofmann, V., Pierrehumbert, J.B., Schütze, H.: Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In: Proceedings of ACL-IJCNLP (2021)

[24] Libovický, J., Helcl, J.: Lexically grounded subword segmentation. In: Proceedings of EMNLP, pp. 7403–7420 (2024)

[25] Di Marco, M., Fraser, A.: Subword segmentation in LLMs: Looking at inflection and consistency. In: Proceedings of EMNLP, pp. 12050–12060 (2024)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- readme.zip