Semi-Automatic Construction and Benchmarking of a Word-Segmented Corpus for Lao Using LLMs and Transformer Models

Ha Nguyen-Tien, Thongphan Palongve, Cuong Nguyen-Quy and Kien Le-Trung
Faculty of Engineering Technology, Hung Vuong University, Phu Tho 350000, Vietnam
E-mail: nguyentienha@hvu.edu.vn, tkvue73@gmail.com, t12345cuong@gmail.com, kiendeptrai06022003@gmail.com

Keywords: Word segmentation, low-resource languages, corpus construction, semi-automatic method, Lao language, NLP

Word segmentation is a fundamental task in Natural Language Processing (NLP), particularly for continuous-script languages such as Lao and Thai, where the absence of spaces between words makes boundary detection highly challenging. In this paper, we present the first publicly released Lao word-segmentation corpus together with a semi-automatic construction pipeline that combines automatic pre-annotation and human verification. Specifically, pre-annotation was generated primarily with GPT-4o (May 2024 snapshot), while a smaller subset was produced using GPT-5o, and all outputs were subsequently corrected through systematic verification by trained native annotators. A senior Lao linguist served as adjudicator to resolve difficult cases and enforce consistency. The final resource comprises 10,000 training sentences and a 1,000-sentence gold-standard test set. To demonstrate utility, we benchmarked Lao word segmentation with an XLM-RoBERTa transformer model, achieving a boundary-level F1 score of 0.75, which surpasses the widely used LaoNLP baseline (0.71). This corpus fills a critical resource gap for Lao and provides a reproducible foundation for future research on word segmentation, POS tagging, machine translation, and other downstream tasks in low-resource language processing.

Povzetek: Članek predstavi prvi javni korpus za laoško segmentacijo besed, zgrajen polavtomatsko, z 10k/1k delitvijo in Transformerji ter demonstrira uporabnost z osnovnim poskusom, kar odpira pot za nadaljnje raziskave.

1 Introduction

Word segmentation is a fundamental task in Natural Language Processing (NLP), especially for continuous-script languages such as Lao and Thai, where words are not separated by spaces or delimiters [21, 22]. This characteristic makes the identification of word boundaries particularly challenging. Accurate segmentation serves as the basis for downstream tasks including machine translation, information retrieval, text summarization, and sentiment analysis [23]. However, the lack of publicly available annotated corpora has severely limited research on Lao NLP [14].

Traditional approaches to word segmentation rule-based or dictionary-based have shown limited success [9, 17]. While rule-based systems provide transparency and flexibility for new words, they require extensive manual effort to create and maintain consistent rules. Dictionary-based approaches can recognize known words effectively but fail with compound terms, dialectal variations, or newly coined words [18]. More recently, machine learning and deep learning methods have demonstrated strong performance in high-resource languages [3, 2], but their heavy reliance on large annotated corpora renders them unsuitable for low-resource settings such as Lao [19, 12].

Although there has been progress in semi-supervised and cross-lingual transfer methods for other low-resource languages [5, 4, 6], there remains a critical gap: Lao still lacks an openly available, high-quality word-segmented corpus. Without such resources, it is difficult for researchers to build, compare, and advance models for Lao NLP [14]. In practice, the absence of a standardized, publicly available corpus means that most prior studies relied on small, privately collected datasets, which vary in domain, size, and annotation quality. This fragmentation prevents fair comparison between models and makes it difficult to reproduce results across research groups. Furthermore, modern neural approaches such as transformer-based models require sufficiently large and consistent annotated data to achieve stable performance; without such data, progress on Lao NLP has remained limited and scattered.

In this paper, we address this gap by introducing the first publicly released word-segmented corpus for Lao. Our corpus was constructed using a semi-automatic method: large language models (LLMs) generated initial annotations, which were then systematically reviewed and corrected by native Lao annotators following annotation guidelines. This approach balances efficiency with quality assurance, enabling the creation of a linguistically consistent dataset at reduced cost. Our contributions are:

1. Proposing a semi-automatic method for constructing

high-quality annotated corpora in low-resource languages, combining large language models for initial annotation with systematic verification by native annotators.

2. Building and releasing the first publicly available corpus of word-segmented Lao sentences, consisting of 10,000 sentences and a gold-standard test set of 1,000 sentences.

3. Benchmarking the dataset with a transformer-based word segmentation model (XLM-R), which achieves state-of-the-art performance on Lao word segmentation, outperforming existing Lao tools by a substantial margin (see Table 5).

Guided by these goals, we pose three research questions to structure the empirical study:

(i) Can LLM-assisted pre-annotation with expert native verification produce a high-quality Lao word-segmentation corpus?

(ii) How does performance scale with corpus size under native-verified supervision?

(iii) To what extent does the resulting corpus improve model generalization relative to existing Lao segmenters?

We answer these in Secs. 3.2–3.4 (pipeline and quality controls), Table 4 (learning curve), and Table 5/Sec. 6.4 (comparisons and analysis).

Our corpus and baseline models are publicly available at `https://github.com/HaHVU/HVULao_NLP`, offering a valuable resource for advancing Lao NLP and supporting future work on other low-resource languages.

2 Related works

Word segmentation and POS tagging in low-resource languages, especially continuous-script languages such as Lao and Thai, remain a challenging task due to limited data and linguistic complexity [21, 22, 14, 24]. Traditional approaches such as rule-based or dictionary-based systems struggle with new words, compound terms, and ambiguous contexts, making them ineffective when used alone [17, 19]. To overcome these limitations, recent research has explored cross-lingual transfer, data augmentation, and semi-automatic methods [2, 5, 11, 12, 13, 6, 8].

Cross-lingual transfer methods have shown promise by leveraging resources from high-resource languages. Conneau et al. [2] introduced XLM-R, a large multilingual model trained on more than 100 languages, which facilitates transfer learning for low-resource languages such as Lao. Similarly, Eskander et al. [5] proposed an unsupervised projection method using parallel corpora (e.g., the Bible) to transfer POS information across languages, demonstrating its potential in extremely low-resource scenarios.

Data augmentation techniques have also been proposed to compensate for the lack of annotated corpora. Shen et al. [12] worked on ancient Chinese, another low-resource continuous-script language, and showed that hybrid tagging schemes and synthetic data generation could significantly enhance both segmentation and tagging accuracy. Plank et al. [13] demonstrated that fine-tuning pre-trained models such as BERT or RoBERTa on small annotated datasets can yield strong results, reducing the need for large-scale resources.

Another research line focuses on semi-automatic or weakly supervised annotation. Sanjeev Kumar et al. [8] further introduced a parallel POS-tagging dataset for Angika, Magahi, Bhojpuri, and Hindi, highlighting the importance of making annotated corpora publicly available for community use. *Comparable challenges and solutions have also been reported for other continuous-script languages in the region such as Khmer, Burmese, and Tibetan.*

For Khmer, recent work has released a 20,000-sentence corpus with manual tokenization and POS tags and established benchmarks from CRF to LSTM–CRF [25]; joint segmentation+POS models based on BiLSTM have also been explored [31]. For Myanmar (Burmese), early segmentation relied on maximum matching and CRF [26, 27], and more recent studies propose joint models (syllable/word segmentation and POS) leveraging BERT [28]. For Tibetan, classical approaches used word-position tagging with CRF [30], while recent neural models perform joint segmentation and POS with BiLSTM/IDCNN/CRF architectures [29, 30]. Taken together, these lines of work highlight a common trend: segmentation-first or joint modeling is necessary in languages with minimal orthographic cues, providing a relevant backdrop for our Lao study.

While the above discussion outlines major methodological directions, Table 1 provides a concise comparison across representative studies, summarizing their language scope, methodological type, dataset availability, and performance. This overview highlights the persistent lack of publicly available Lao resources and justifies our focus on constructing a fully open, native-verified corpus.

Beyond word segmentation and POS tagging, several recent studies have also emphasized the importance of constructing reliable corpora and domain-specific resources for advancing NLP. For example, Hu et al. [32] developed a clinical corpus for characterizing pituitary adenomas and demonstrated its application in large language models. Ahmad et al. [33] designed and evaluated a hate speech detection corpus for Arabic. Shao et al. [34] proposed an integrated NLP method for text mining and visualization of underground engineering reports. Although focusing on different application domains, these works share with our study the same objective of building high-quality linguistic resources that enable modern NLP research.

Table 1: Related work on segmentation/POS for low-resource continuous-script languages: methodology, key method/model, dataset availability, benefits, and drawbacks.

Ref.	**Language(s)**	**Methodology / Key method**	**Dataset availability**	**Benefits**	**Drawbacks**
[21, 22]	Thai	Data-driven word segmentation	Not public (task-specific)	data-based learning	Domain dependence; limited generalization
[14]	Lao	Dictionary-based segmentation	Open dictionaries; Y-Cup-2020 corpus	Open-source toolkit	no deep learning or contextual modeling
[25, 31]	Khmer	BiLSTM (char) joint seg + POS	Khmer POS (12k)	Unified; 97.1% seg. acc.; simple	Small dataset; no emb.; mono Khmer
[12, 32, 34]	Chinese	Hybrid tagging + *data augmentation*/CCNLP platform/ BERT + BiLSTM + CRF	not public	Boosts seg/POS with synthetic data	Dependent on pseudo-data quality; domain shift
[26, 27, 28]	Burmese	CRF–SVM–LM compare; CRF boundary; BERT joint tagger	non-public	joint reduces errors; 1st neural Burmese WSG–POS	Limited domain; inconsistent annotation
[30, 29]	Tibetan	BiLSTM+IDCNN+CRF joint; 6-tag CRF + post-processing	non-public	OOV-robust; effective joint model	Feature-based; needs larger corpus
[17, 19]	Slovene/ Vietnamese	lexicon-based/Hybrid	available for research	handles overlap ambiguities effectively	depends on lexicon completeness
[8]	Angika, Magahi, Bhojpuri, Hindi	Min.-supervised morph. seg. (Morfessor / AG / CRF)	4-lang wordlists; annotated/unannotated	Comparative; error-analytical; low-resource guidance	No novel model; 4-lang scope only
[3]	English, Estonian, Finnish, Turkish	MuRIL-Hi + Look-back	708-sent. UD-parallel, public	Token fix; strong zero-shot	Small / domain-limited; relies on Hindi transfer
[2]	100+ (incl. Lao)	Transfer learning with **XLM-R**	Code/models, CC-100 data	SOTA on XNLI/MLQA/NER	extremely compute-intensive
[5]	12 target + 6 source	annotation projection + BiLSTM	Parallel corpora	Unsupervised; robust multi-source transfer	Domain bias; limited lexical/morph coverage
[11]	104 languages	mBERT transfer (fine-tuning)	Models public; Wikipedia (104 langs)	Good POS transfer with small labels	Inconsistent for continuous scripts; needs gold data
[13]	22 languages	BiLSTM + W-C Emb + LOGFREQ Loss	v1.2, WSJ public	OOV-robust; low-data generalization	Label-noise sensitive; unstable on small datasets
[6]	17 low-resource + 15 high-resource	GLP = GNN Label Propagation + Self-Learning	Parallel Bible, UD v2.10 public	Unsupervised POS transfer; works for unseen langs	High compute; Bible-domain bias
[15]	101 languages	Subword tokenizer (SentencePiece/BPE) as heuristic segmentation	Corpora public (Flores)	Easy to apply cross-lingually	Heuristic splits; not true Lao word boundaries
This work	**Lao**	**LLM-assisted pre-annotation + native adjudication; XLM-R fine-tuned**	**Fully released (10k train + 1k gold test)**	**High-quality, reproducible Lao corpus; SOTA Lao segmentation**	**Current domains: news/official; social/dialogue future work**

Despite these advances, Lao still lacks an openly available, high-quality corpus for word segmentation [14, 15]. Most existing studies rely on transfer or augmentation but do not address the fundamental bottleneck of corpus construction. Our work differs from prior approaches by proposing a semi-automatic method that not only leverages large language models for initial annotation, but also ensures linguistic quality through systematic verification by native Lao annotators. This combination enables us to build and release the first public word-segmented corpus for Lao, filling a critical resource gap and providing a reproducible foundation for future low-resource NLP research.

3 Our proposed method

Our goal is to efficiently construct a high-quality, publicly released word-segmented corpus for Lao using a semi-automatic pipeline that combines large language models (LLMs) with systematic human verification. The pipeline is illustrated in Figure 1 and consists of three main stages:

1. **Pre-annotation with LLMs (No1 → No2).** Raw Lao sentences collected from reliable sources are first pre-annotated by an LLM (e.g., GPT-5o) to produce raw word-segmented candidates.

2. **Expert verification by native annotators (No2 → No3).** Native Lao annotators review and correct the pre-annotations following written guidelines, resolving boundary ambiguities, compound words, and edge cases. The verified outputs are consolidated into the high-quality word-segmented corpus.

3. **Model bootstrapping.** The verified corpus (No3) is then used to fine-tune a transformer-based word-segmentation model, reducing future reliance on LLMs and lowering annotation cost.

This design explicitly balances efficiency and quality: LLMs drastically reduce manual effort, while linguist verification ensures internal consistency and prevents systematic LLM errors from propagating.

3.1 Collecting Lao monolingual sentences

The first stage of our semi-automatic pipeline involves the collection of naturally occurring Lao sentences to construct a representative monolingual corpus for word segmentation. To ensure linguistic reliability and coverage, we deliberately targeted trustworthy sources that follow grammatical conventions and provide domain diversity. Specifically, we gathered texts from official government websites, national news portals, academic institutions, and legal repositories. These domains cover a broad range of topics including politics, education, law, culture, and daily news thereby contributing to both lexical richness and structural variety, which are essential for developing robust NLP resources.

All collected texts underwent a standardized pre-processing procedure. This included removing duplicated sentences, discarding incomplete or noisy fragments, and normalizing formatting inconsistencies such as irregular punctuation or encoding artifacts. Through this filtering process, we ensured that only linguistically sound and representative sentences were retained for subsequent processing.

As a result, we compiled a sufficiently large dataset to support both training and evaluation. Specifically, we constructed a gold-standard test set of 1,000 sentences for reliable benchmarking, and a high-quality training corpus of 10,000 sentences for model development. Table 2 summarizes the primary sources from which the monolingual Lao data were collected.

3.2 LLM-based pre-annotation

In this stage, we leverage a large language model to generate initial word-segmentation candidates for Lao sentences. The model is instructed to output only segmented text in a strict, machine-checkable format, without any additional linguistic annotations. This design minimizes ambiguity in the output and simplifies subsequent verification by human annotators.

To increase the robustness of LLM-generated segmentations in a low-resource setting, we adopt the following strategies:

– **Task-specific prompting.** The prompt provides explicit instructions to segment Lao sentences into word tokens, along with format requirements that enforce one token per unit separated by spaces or line breaks.

– **Few-shot demonstrations.** The prompt includes short Lao examples that illustrate typical segmentation challenges, such as reduplications, clitics, compound words, and punctuation, thereby guiding the model toward consistent boundary decisions.

– **Error-driven refinement.** Prompts are iteratively refined based on systematic error analyses during pilot experiments. Frequent issues, such as over-segmentation of function words or under-segmentation around punctuation, are explicitly highlighted and corrected through improved instructions and examples.

For reproducibility, each pre-annotated sentence is stored together with its original unsegmented form. This ensures provenance tracking and allows researchers to replicate or refine the pre-annotation process in future studies.

In practice, the majority of pre-annotations were generated with GPT-4o (OpenAI, API snapshot May 2024). A small portion of pilot data was also produced using

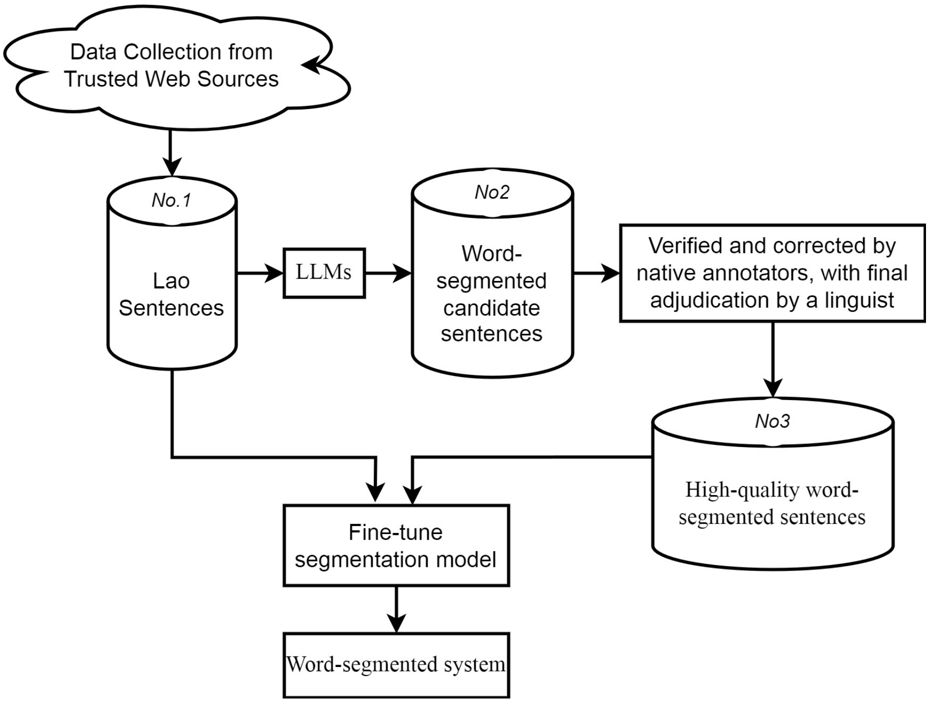
Data Collection from Trusted Web Sources
No.1
Lao Sentences
LLMs
No2
Word-segmented candidate sentences
Verified and corrected by native annotators, with final adjudication by a linguist
No3
High-quality word-segmented sentences
Fine-tune segmentation model
Word-segmented system

Figure 1: Semi-automatic pipeline for constructing the Lao word-segmented corpus

Table 2: Domain sources and sentence distribution in the Lao monolingual corpus

No.	Source / Domain	Sentences	Proportion (%)
1	National Assembly of Laos (na.gov.la)	1,240	12.4
2	Lao News Agency – KPL (kpl.gov.la)	2,130	21.3
3	Ministry of Education and Sports (moes.edu.la)	920	9.2
4	Lao Official Gazette (laogov.la)	870	8.7
5	National University of Laos (nuol.edu.la)	960	9.6
6	Vientiane Times (vientianetimes.org.la)	1,980	19.8
7	VOV Lao (vovworld.vn/lo-LA.vov)	1,900	19.0
Total	**7 trusted sources**	**10,000**	**100.0**

the subsequently released GPT-5o, demonstrating that the pipeline is model-agnostic and can flexibly adopt future LLMs.

We configured GPT-4o and GPT-5o with the following decoding parameters: *temperature = 1.0, top_p = 1.0, max_tokens = 512, frequency_penalty = 0, presence_penalty = 0, stop = none.* This fixed configuration, together with the strict prompt design, ensured stable and format-consistent outputs.

For reproducibility, we documented the model identifier, snapshot date, and prompt version used in the pre-annotation process.

An excerpt of the instruction prompt is presented below:

You are a Lao word segmentation assistant.
TASK: Insert spaces between Lao word tokens.
RULES:
1) Output only the segmented sentence.
2) One token per unit, separated by spaces.
3) Punctuation marks and particles are separate tokens.

Input:
ທ່ານຟ້າມບິນມິນເນັ້ນໜັກວ່າແຫຼ່ງທຶນຂອງທະນາຄານພັດທະນາອາຊີໄດ້ນຳໃຊ້ຢ່າງມີປະສິດທິຜົນ,

Output:
ທ່ານ ຟ້າມບິນມິນ ເນັ້ນໜັກ ວ່າ ແຫຼ່ງທຶນ ຂອງ ທະນາຄານ ພັດທະນາ ອາຊີ ໄດ້ ນຳໃຊ້ ຢ່າງ ມີ ປະສິດທິຜົນ ,

(Mr. Pham Binh Minh emphasized that the Asian Development Bank's funds have been used effectively.)

After generating automatic pre-annotations as candidate segmentations, the next step involves manual verification to ensure linguistic accuracy and consistency.

3.3 Verification by native annotators

All candidate segmentations produced by the LLM are systematically verified by trained native Lao annotators through a custom web-based interface. This interface visualizes differences between the LLM output and the corrected version, allowing annotators to focus on token boundary modifications rather than re-annotating entire sentences.

Annotators strictly follow a written segmentation guideline that was developed to ensure consistency and reproducibility across annotators. The guideline specifies detailed treatments for common and challenging cases, including:

- Handling of compound words, multi-word expressions, numerals, dates, and named entities;
- Disambiguation heuristics for function words, particles, and ambiguous morphemes;
- Normalization rules for punctuation and mixed-script tokens (e.g., Latin characters or digits).

To balance efficiency and quality, edits are intentionally minimal and surgical: annotators correct only where necessary (fix-only), thereby maintaining annotation throughput while ensuring internal consistency. When disagreements arise, cases are escalated to a native linguist, who adjudicates and finalizes the decision.

Annotators underwent a short training phase on a pilot set with iterative feedback from the senior linguist to ensure consistent application of the guidelines. To assess annotation consistency, we randomly selected 5% (500 sentences) for double annotation. The inter-annotator segmentation F1 was 0.80, indicating substantial agreement and reliable boundary consistency.

In practice, the annotation team consisted of four trained native annotators and one native linguist, working over a period of eight months to complete the verification process. This sustained effort highlights both the difficulty of Lao word segmentation and the scale of our contribution. In particular, the 1,000-sentence test set was carefully constructed and validated by the native linguist to ensure its reliability as a gold-standard benchmark, providing a trusted basis for evaluating segmentation models and supporting reproducible research.

3.4 Joint pre-annotation and adjudication (segmentation + POS)

For efficiency, we obtain a *joint* LLM pre-annotation that includes both word boundaries and POS labels. Trained native annotators then correct segmentation and POS in a single pass using a diff-based interface; disagreements are adjudicated by a native linguist. The released dataset therefore contains (i) a gold word-segmentation layer and (ii) an aligned gold POS layer. To keep the scope focused, this paper evaluates models only for word segmentation, while the POS layer is released as an additional resource and its evaluation is left for future work.

4 Corpus statistics and analysis

To assess the coverage and linguistic diversity of our resource, we performed a systematic analysis of both the training corpus (10,000 sentences) and the gold-standard test set (1,000 sentences). The statistics are summarized in Table 3.

The corpus exhibits solid lexical coverage, with 13,339 unique word types in the 10k training set and 2,901 in the 1k test set. The average sentence length is 19.37 tokens in training and 15.73 tokens in test, which is consistent with written Lao drawn from news, academic, and legal sources. Importantly, **80.1%** of the word types in the test set also appear in the training vocabulary, indicating substantial lexical overlap and making the resource suitable for benchmarking segmentation models.

Figure 2 reports the sentence length distribution for the 10k training corpus. The distribution is right-skewed with a clear mode at 10–14 tokens. The bulk of sentences lie between 10 and 29 tokens, a smaller but noticeable portion are short (under 10 tokens), and long sentences become progressively rarer; only a small tail contains 40+ tokens. This shape reflects the structural variety of Lao, from short declaratives to longer multi-clause sentences, and its implications for corpus utility are worth noting.

In particular, the majority of sentences are of medium length, which provides abundant context for learning boundary patterns. The concentration around 10–29 tokens suggests that the corpus predominantly captures expository and declarative structures, typical of news and formal prose. By contrast, the sparsity of very long sentences indicates that highly nested or multi-clause constructions are under-represented, which may limit generalization in those cases. Thus, while the training set is well-suited for developing segmentation models in standard written contexts, future extensions could benefit from incorporating more long-form texts to cover complex syntactic patterns.

Complementing the training set, the gold-standard test set exhibits a similar right-skewed profile with a mode at 10 - 14 tokens and an average of 15.73 tokens per sentence. Most sentences fall between 5 and 29 tokens ($\approx$ **94.1**%), with the majority concentrated in 10–29 tokens ($\approx$**77.6%**), short sentences under 10 tokens account for $\approx$ **17.3**%, sentences of 30–39 tokens make up $\approx$ **4.3**%, and very long sentences ($\geq$ 40 tokens) are rare ($\approx$ **0.5**%). This mirrors the structural variety observed in Lao prose while keeping a long but light tail of complex sentences (see Figure 3), and its distribution offers further insights into evaluation conditions.

Like the training set, the majority of test sentences are of medium length, which provides abundant context for assessing segmentation accuracy in standard prose. The

Table 3: Statistics of the Lao word-segmented corpus

Statistic	**Training Corpus (10k)**	**Test Set (1k)**
Number of sentences	10,000	1,000
Total tokens	193,714	15,730
Mean sentence length	19.37	15.73
Average characters per token	4.32	4.26
Unique word types	13,339	2,901
Vocabulary coverage (test vs. train)	80.1%	

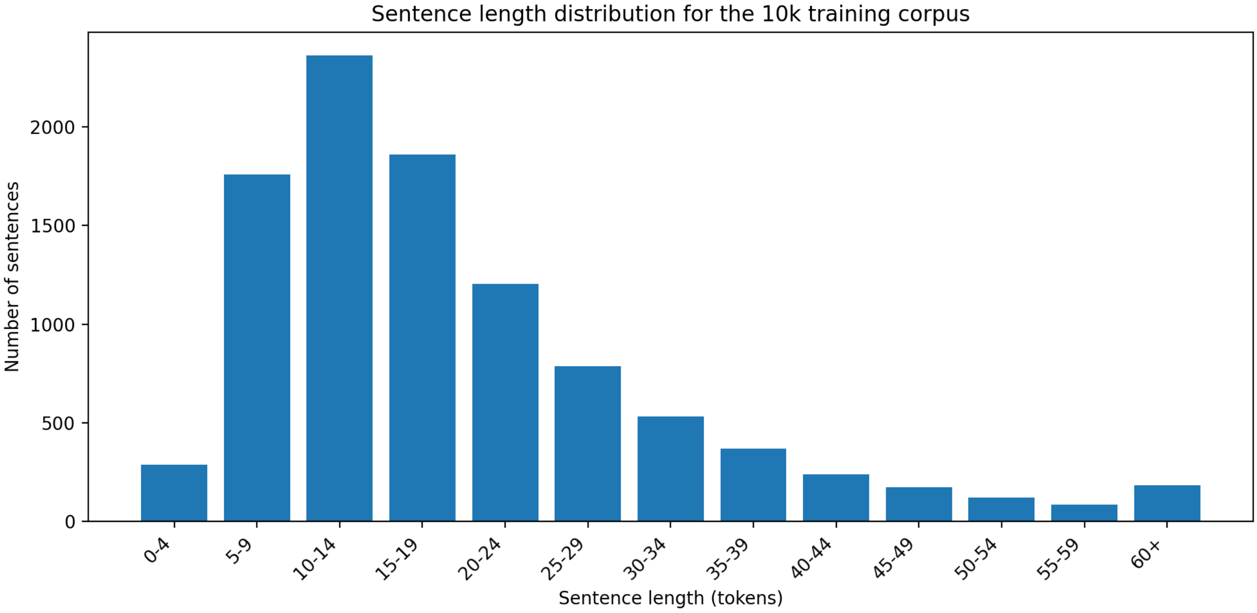

Figure 2: Sentence length distribution (in tokens) for the 10k training corpus

noticeable share of short sentences ensures coverage of cases dominated by function words and particles, while the scarcity of very long sentences means that robustness on deeply nested structures should be interpreted with caution. Overall, the test distribution closely aligns with the training profile, which minimizes domain shift and supports reliable benchmarking, but also highlights the need for future expansions with more conversational and complex sentence types.

5 Corpus linguistic characteristics

A qualitative inspection of the corpus highlights several key linguistic features:

- Compound words and reduplication: Lao frequently forms compound nouns and uses reduplication for emphasis or plurality, both of which challenge segmentation.
- Function words and particles: Common particles (e.g., negation, aspect markers, discourse particles) are often short and can be easily under-segmented.
- Named entities and numerals: Mixed-script tokens (Lao with Latin characters or digits) appear frequently in news and official texts, requiring consistent annotation.
- Ambiguity in boundaries: Certain morphemes serve multiple syntactic functions depending on context, demanding careful disambiguation during annotation.

These characteristics underline the importance of having a carefully validated gold-standard test set. As described in Section 4, the 1,000-sentence test set was meticulously constructed and adjudicated by a native linguist, ensuring high reliability and reproducibility for future benchmarking.

Overall, the corpus provides not only scale but also linguistic variety, making it a valuable resource for training robust segmentation models and studying Lao as a low-resource continuous-script language.

POS layer: Beyond boundaries, the corpus includes a gold POS layer vetted by native annotators and aligned to the released tokenization. We provide the tagset, labeling rules, and examples of difficult cases to ensure consistent downstream use; POS modeling is left to future work.

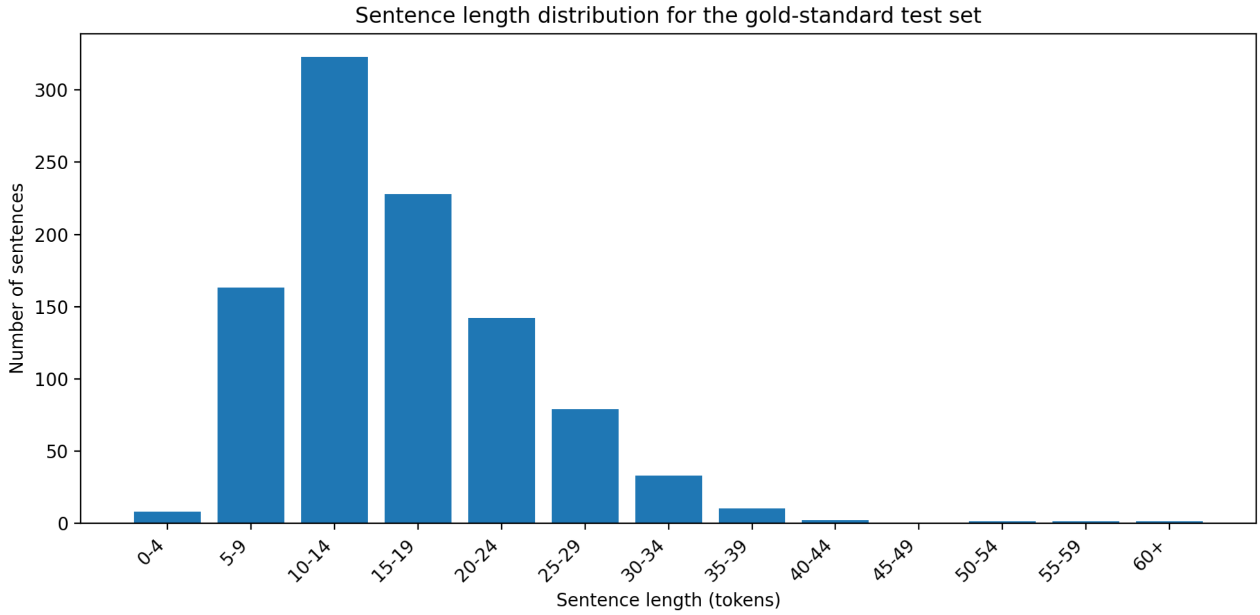

Figure 3: Sentence length distribution (in tokens) for the gold-standard test set

6 Experiments

6.1 Task, data splits, and metrics

Our goal is to verify that the proposed corpus enables accurate *Lao word segmentation*. We treat segmentation as token classification with a **BIO** scheme over subword tokens. Given the tokenizer's offset mapping, we label the first subword of a word as `B-WORD`, subsequent subwords as `I-WORD`, and non-text/padding as `O`; gold boundaries are recovered from the labels.

We use the **10k** verified sentences for training/validation with a **15%** random validation split (seed = 42). The **1k** native-validated set is held out strictly for final testing. Evaluation is conducted at the *boundary level*: a predicted boundary is correct *iff* it matches a gold word boundary. We report **Precision**, **Recall**, and **F1**.

6.2 Model and training setup

We fine-tune **XLM-RoBERTa-base** [2] using the Hugging Face transformers stack. Training uses **AdamW** with learning rate 2×10^{-5}, weight decay 0.02, sequence length 128, batch size 4 per device with **gradient accumulation** = 4, and **10** epochs. We enable `fp16` mixed precision and a warmup ratio of **0.1**. Models are evaluated every epoch and the best checkpoint is selected by validation loss with **early stopping** (patience = 5, at most two checkpoints retained).

Computing environment All experiments were conducted on a dedicated server running Ubuntu 20.04.6 LTS (Focal Fossa), equipped with an NVIDIA GeForce RTX 3090 GPU (24 GB VRAM). The implementation used PyTorch and Hugging Face Transformers, following the setup described above (XLM-RoBERTa-base backbone, sequence length 128, batch size 4 per device with gradient accumulation 4, AdamW with learning rate 2×10^{-5}, weight decay 0.02, warmup ratio 0.1, 10 epochs, fp16 enabled, evaluation/saving each epoch, and early stopping with patience 5). On this setup, each training run required about 40 minutes.

6.3 Learning curve

To examine data efficiency, we fine-tune with {2k, 4k, 6k, 8k, 10k} training sentences while keeping the test set fixed. Results on the 1k gold test set are shown in Table 4.

Table 4: Learning-curve results on the 1k gold test set (boundary-level). Values are **mean** ± **SD** over five random seeds (13, 21, 42, 87, 123).

Training Data	Precision	Recall	F1 score
2,000	0.64 ± 0.01	0.46 ± 0.02	0.53 ± 0.01
4,000	0.73 ± 0.02	0.64 ± 0.01	0.68 ± 0.01
6,000	0.69 ± 0.01	0.75 ± 0.01	0.72 ± 0.01
8,000	0.70 ± 0.01	0.75 ± 0.01	0.73 ± 0.01
10,000	0.76 ± 0.01	0.74 ± 0.01	**0.75** ± 0.01

Observations. (i) Performance increases steadily with more training data; gains start to taper after $6 - 8$k sentences, suggesting diminishing returns once the model has seen sufficient boundary patterns. (ii) Improvements at larger sizes indicate that *native-verified* consistency benefits generalization as much as scale.

To assess robustness, we fine-tuned five independent runs with different random seeds (13, 21, 42, 87, 123).

Table 4 reports the mean ± standard deviation (SD) of each metric. The extremely small SD values (≤ 0.02) demonstrate that the model's learning behavior is stable and not sensitive to random initialization, confirming the statistical reliability of the observed trend.

6.4 Comparison with existing Lao segmenters

We compare our model against five baselines: a machine learning–based sequence labeling model (**CRF++**[1]), a widely used Lao segmenter (**LaoNLP**[2]), a subword-based tokenizer from the **Flores**-101 benchmark (**Flores**[3]), a multilingual transformer baseline (**mBERT**[4]), and a Thai Word Segmentation using SentencePiece model (**Thai-SP**[5]). Table 5 reports boundary-level scores on the same 1k gold test set. [14, 15, 3, 11, 7, 16].

Table 5: Comparison on the 1k gold test set (boundary-level)

Model	Precision	Recall	F1 Score
CRF++	0.15	0.04	0.06
LaoNLP	0.71	0.71	0.71
Flores	0.56	0.45	0.50
Thai-SP	0.34	0.10	0.15
mBERT	0.33	0.11	0.16
Our model	**0.76**	**0.74**	**0.75**

For completeness, both mBERT and Thai-SP were evaluated under the same boundary-level setting on the 1k gold test set. Their weak performance (mBERT: F1 = 0.16; Thai-SP: F1 = 0.15) indicates that neither multilingual transformers nor related-language tokenizers transfer effectively to Lao without in-language supervision.

The weak performance of CRF++ (F1 = 0.06) and mBERT (F1 = 0.16) further confirms the inherent difficulty of Lao segmentation and the necessity of a high-quality corpus. These baselines serve not merely as points of comparison but as evidence that our corpus effectively exposes the limits of existing methods, traditional statistical models lack contextual understanding, while generic multilingual transformers and cross-lingual tokenizers struggle with continuous scripts and low-resource settings. Our fine-tuned XLM-R model outperforms all baselines across every metric. Compared with CRF++ and mBERT, it achieves substantially higher precision and recall, confirming that contextualized representations capture boundary regularities that traditional statistical and generic multilingual models fail to learn. When evaluated against the off-the-shelf LaoNLP segmenter under its default configuration, XLM-R demonstrates a markedly better balance between precision and recall, effectively reducing both over-segmentation and under-segmentation. The weaker performances of Flores and Thai-SP further underscore the importance of language-specific supervision and in-language adaptation for Lao.

7 Discussion

Our experiments address a central question: *does the proposed corpus enable accurate and reproducible Lao word segmentation?* The results support an affirmative answer and lead to several observations.

Data quality and coverage: The corpus balances scale and linguistic variety: the 10,000-sentence training set contains 13,339 word types and the 1,000-sentence test set contains 2,901 types, with 80.1% test types observed in training. Length profiles for train and test are similarly right-skewed (mode at 10–14 tokens), indicating that the held-out test is representative of the training distribution. Native-linguist adjudication enforces internal consistency, which is crucial for boundary-level supervision.

Effect of supervision volume: The learning curve in Table 4 shows steady gains as the number of supervised sentences grows from 2k to 10k, culminating in an F1 score of 0.75. Improvements taper after 6-8k sentences, suggesting diminishing returns once the model has encountered a sufficiently broad range of boundary patterns. This highlights that *label quality and consistency* are at least as important as raw quantity in low-resource settings.

Across five random seeds, the variance of Precision, Recall, and F1 was minimal (SD ≤ 0.02), indicating high experimental reproducibility. This confirms that performance differences across training-set sizes are statistically consistent rather than artifacts of random initialization.

Comparison to existing systems: On the same 1k gold test set and under boundary-level evaluation, our fine-tuned XLM-R model outperforms all baselines, including LaoNLP, Flores, Thai-SP, CRF++, and mBERT (Table 5). Gains are balanced in Precision and Recall, indicating fewer over-segmentations and fewer missed boundaries. In summary, contextualized representations trained on a native-verified corpus offer measurable advantages over simple subword-based segmentation heuristics (e.g., SentencePiece/BPE from the Flores pipeline) and over generic multilingual tokenizers.

Beyond absolute performance, the inclusion of CRF++, mBERT, and Thai-SP baselines provides a broader empirical context for assessing segmentation difficulty. The extremely low F1 of CRF++ (0.06) demonstrates that surface-level or feature-based models are unable to capture the morphosyntactic regularities of Lao, while mBERT (F1 = 0.16) and Thai-SP (F1 = 0.15, evaluated zero-shot on Lao) highlight that even large multilingual

[1] https://taku910.github.io/crfpp/

[2] https://github.com/wannaphong/LaoNLP

[3] https://github.com/facebookresearch/flores

[4] https://huggingface.co/bert-base-multilingual-cased

[5] https://github.com/wannaphong/thai-word-segmentation-sentencepiece

Table 6: Breakdown of segmentation error types on the 1k gold test set. Percentages are measured over all boundary mismatches.

Error Type	Proportion	Typical Example
Over-segmentation	**7.0%**	*Reference:* ບຸນປີໃໝ່ລາວ (‘Lao New Year’) *Predicted:* ບຸນ ປີ ໃໝ່ ລາວ
Under-segmentation	**6.0%**	*Reference:* ທະນາຄານ ພັດທະນາ ອາຊີ (‘Asian Development Bank’) *Predicted:* ທະນາຄານພັດທະນາອາຊີ
Total Error Rate	**13.0%**	

or related-language models fail to transfer effectively to a continuous-script, low-resource language. These observations reinforce that Lao presents a challenging test case for multilingual NLP, and that language-specific supervision remains indispensable despite recent advances in universal representation learning. Consequently, our corpus not only enables reproducible benchmarking but also establishes a clear baseline landscape for future research on more systematic transfer learning and cross-lingual adaptation.

Typical errors and quantitative analysis: A quantitative error breakdown was conducted on the 1k gold-standard test set to better characterize residual segmentation issues. Boundary mismatches were categorized into two main types: *over-segmentation* (spurious boundaries predicted inside a gold token) and *under-segmentation* (missing boundaries between adjacent gold tokens). Overall, the model achieved an **F1 score of 0.75**, with a small Precision–Recall gap (0.76 vs. 0.74) indicating a mild tendency toward predicting extra splits. Table 6 summarizes the proportion of each error type, showing that over-segmentation slightly dominates (7.0%) over under-segmentation (6.0%), yielding an overall boundary error rate of 13%.

Representative examples: Closer inspection of model outputs reveals systematic error patterns that reflect the morphosyntactic nature of Lao. In ທະນາຄານພັດທະນາອາຊີ (‘Asian Development Bank’), the model produced a single token instead of the correct segmentation ທະນາຄານ ພັດທະນາ ອາຊີ, illustrating typical *under-segmentation* of multi-word named entities. Conversely, in ບຸນປີໃໝ່ລາວ (‘Lao New Year’), the model split the lexicalized compound into four separate words (ບຸນ ປີ ໃໝ່ ລາວ), a clear instance of *over-segmentation*. Another recurrent boundary ambiguity involves short function words and clitics, such as in ໄປບ່ອນໃດ (‘go where’), where the model occasionally inserted an unnecessary split (ໄປ ບ່ອນໃດ).

These patterns are linguistically interpretable rather than random. Over-segmentation tends to affect frequent compounds and reduplicative structures, while under-segmentation mainly occurs with named entities or tightly bound multi-word expressions. Together with the low variance observed across five random seeds (SD $\leq$ 0.02, see Table 4), this analysis confirms that segmentation errors are systematic and stable across experimental runs. It also highlights specific areas where annotation guidelines and model post-processing can be refined in future work.

Limitations and threats to validity: Although the corpus spans multiple public domains, news/legal prose remains over-represented; extending to conversational, social-media, and technical registers would improve robustness. As a result, informal or interactive genres (e.g., everyday conversations, social media) are under-represented. Including these domains in future releases will enhance coverage and generalizability. Adjudication maximizes final label quality but limits conventional reporting of inter-annotator agreement; to address this, we plan to release a double-annotated subset to estimate IAA and enable disagreement-aware training. While the experiments in this paper focus on *word segmentation*, we *release* a gold POS layer produced via the same joint annotation–adjudication pipeline; benchmarking POS is deferred to future work to keep the evaluation centered on a single task. This clarification makes explicit that the POS layer is provided as a complementary resource, but its evaluation is intentionally outside the scope of this paper. Finally, LLM-assisted pre-annotation may introduce systematic biases and our boundary supervision relies on tokenizer offset mapping; native-linguist adjudication mitigates these risks but residual bias and tokenizer sensitivity cannot be fully excluded.

8 Conclusion and future work

We introduced the first publicly released corpus for *Lao word segmentation* and a practical semi-automatic pipeline that couples LLM pre-annotation with native-annotators' adjudication. The resource comprises a 10k training set and a 1k gold test set with clear guidelines and provenance. Corpus statistics show solid lexical coverage and matched length profiles across train and test, and a transformer baseline (XLM-R base) fine-tuned on our data attains **F1=0.75**, surpassing all evaluated baselines, including CRF++, LaoNLP, Flores, mBERT, and Thai-SP. These results demonstrate that efficiently obtained, *native-verified* supervision is sufficient to enable accurate and reproducible segmentation for a continuous-script, low-resource language.

The comparative results with CRF++, mBERT, and Thai-SP further illustrate that Lao segmentation remains a non-trivial challenge even for modern multilingual and related-language models, thereby underscoring the benchmark value of the proposed corpus. Preliminary experiments with Thai-SP highlight the difficulty of cross-lingual transfer, suggesting that future extensions could explore more systematic fine-tuning or adapter-based transfer from typologically related languages such as Thai or Khmer to further test the dataset's generalization potential.

Future work: We will (i) **scale and diversify** the corpus beyond 10k sentences with broader domains, including conversational, forum, and social-media texts, to reduce the current bias toward news and official documents, (ii) release a **double-annotated subset** to enable inter-annotator agreement measurement and disagreement-aware learning, (iii) design **active/uncertainty-driven annotation** workflows to reduce human effort while preserving quality, (iv) provide **phenomenon-targeted test suites** (compounds, reduplication, mixed-script tokens) for focused evaluation, and (v) investigate stronger baselines (e.g., XLM-R+CRF, character-aware models, self-training/cross-lingual transfer) and, in a separate track, extend annotations to **POS** and named entities.

Ethical considerations

All data used in this study were collected from publicly available government and media websites that explicitly allow text reuse for research and non-commercial purposes. No personally identifiable or confidential information was included. Pre-annotation with GPT-based large language models (LLMs) was conducted offline, and the outputs were manually verified and corrected by native linguists to mitigate potential biases or hallucinations introduced by the models. The authors acknowledge that LLM-generated annotations may reflect biases inherent in their training data; however, all final corpus releases were carefully reviewed to ensure neutrality, linguistic accuracy, and cultural appropriateness. The dataset and scripts are shared under an open license to encourage transparency and responsible reuse in future research.

Acknowledgements

The authors would like to thank the anonymous reviewers and editors for their insightful comments and constructive suggestions that helped improve the quality of this paper. The authors also acknowledge Hung Vuong University, Phu Tho, for its continuous academic and research support.

Funding

This work was supported by Hung Vuong University, Phu Tho, under the project grant number 25/2024/KHCN (HV25.2024).

Appendix A. Dataset description and reproducibility

A.1 Release contents

We provide two dataset splits with strictly preserved line ordering (UTF-8 NFC).

- **Test** (`Datatest1k/`): `testorgin1000.txt` (raw; one sentence per line), `testsegsent_1000.txt` (word-segmented), and `testtag1k.json` (word-segmented + POS-tagged).

- **Train** (`Datatrain10k/`): `10korinsentecces.txt` (raw), `10ksegmented.txt` (word-segmented), and `10kcorpustag.json` (word-segmented + POS-tagged).

All raw files are encoded in UTF-8 (NFC). Each JSON file stores every sentence as an array of token objects {`word`, `label`}, aligned 1-to-1 with the corresponding raw and segmented lines.

A.2 Schema example (from testtag1k.json)

```
[
  [
    {"word": "ທ່ານ", "label": "N"},
    {"word": "ຄິດ", "label": "V"},
    {"word": "ວ່າ", "label": "COJ"},
    {"word": "ລາວ", "label": "N"},
    {"word": "ຈະ", "label": "PRA"},
    {"word": "ມາ", "label": "V"},
    {"word": ".", "label": "PUNCT"}
  ],
  [
    {"word": "ທ່ານ", "label": "N"},
    {"word": "ຮູ້", "label": "V"},
    {"word": "ບໍ", "label": "IAQ"},
    {"word": "ລາວ", "label": "N"},
    {"word": "ມາ", "label": "V"},
    {"word": ".", "label": "PUNCT"}
  ]
]
```

A.3 Alignment invariants and examples

Alignment invariants:

- Raw line $i \leftrightarrow$ segmented line $i \leftrightarrow$ JSON entry i.

- Segmented sentence = whitespace join of all `word` fields in JSON[i].

- Order is stable; no shuffling between files.

- Encoding: UTF-8 NFC; zero-width and control chars removed.

Example of alignment:

ທ່ານຄິດວ່າລາວຈະມາ. ⇒ ທ່ານ(N) + ຄິດ(V) + ວ່າ(COJ) + ລາວ(N) + ຈະ(PRA) + ມາ(V) + .(PUNCT)

ທ່ານຮູ້ບໍລາວມາ□ ⇒ ທ່ານ(N) + ຮູ້(V) + ບໍ(IAQ) + ລາວ(N) + ມາ(V) + .(PUNCT)

A.4 POS tagset (summary)

We use a compact Lao-oriented POS inventory adapted from UD conventions. Each function word or particle is treated as a separate token to ensure consistent POS alignment.

Example:

ຈະ ພັດທະນາ ⇒ ຈະ (PRA) + ພັດທະນາ (V)

A.5 Reproducibility pointers

We provide a command-line interface segmentation tool (GPU-ready) trained on the 10 k split:

```
python3 scripts/segment_lao.py -i <input_file>

-o <output_file>
```

Arguments

-i, –input: Path to the input file (required)

-o, –output: Path to the output file (required)

All dependency versions and training scripts are pinned in the repository (see `requirements.txt` and `scripts/`).

A.6 Environment and dependencies

All experiments were conducted on a workstation with NVIDIA RTX 3090 (24 GB VRAM). Reproducible environment versions:

```
torch==2.3.0
transformers==4.53.0
datasets==2.21.0
accelerate==0.33.0
python>=3.7
```

The tool automatically uses GPU if available and falls back to CPU otherwise.

A.7 Fine-tuning configuration and reproducibility

The corpus was used to fine-tune `xlm-roberta-base` using `scripts/FineTuneSegLaowithseeds.py`.

- Tokenizer: `XLMRobertaTokenizer`
- Max sequence length: 128
- Batch size 4 (per device), gradient accumulation 4
- Learning rate 2×10^{-5}, weight decay 0.02
- Warm-up ratio 0.1, epochs 10, FP16 enabled
- Random seeds $\{13, 21, 42, 87, 123\}$

Outputs: `results_multi_seed/` (models), `logs_multi_seed/` (training logs), `eval_multi_seed/` (predictions).

Mean ± std of F1 across seeds = 0.75 ± 0.01 (as reported in the paper).

Appendix B. Segmentation guidelines

Unit of annotation: Segmentation is performed at the *word* level. Tokens are the minimal orthographic units that carry lexical or grammatical meaning. We adopt a "fix-only" principle during verification: correct only what is necessary to ensure consistent gold-standard boundaries.

General principles

- **Minimality & consistency**: prefer the smallest segmentation consistent across the corpus.
- **Context sensitivity**: ambiguous cases are resolved by sentential context, not frequency alone.
- **Native-verified rules**: disagreements are adjudicated by a trained native linguist following written guidelines.

Boundary rules

1. **Function words and particles**: always separate as individual tokens. e.g., preverbal auxiliaries ຈະ, ໄດ້; discourse particles ກໍ່, ເດີ.
2. **Compounds and multi-word expressions**:
   - If orthographically fused and lexicalized, keep as one token (e.g., ນຳໃຊ້).
   - Otherwise, segment by component words (e.g., ທະນາຄານ ພັດທະນາ ອາຊີ).
   - **Decision test**: if a component can be modified/replaced independently, segment it.
3. **Affix-like morphemes and bound particles**: Productive morphemes such as ບໍ່ (NEG) or nominalizer ກຳ are separated when used syntactically, but remain fused in lexicalized forms (e.g., ກຳລັງ "energy" → one token).
4. **Reduplication**: each reduplicant is a separate token (e.g., ແທ້ ໆ, ໄວ ໆ).
5. **Mixed-script sequences**: Latin letters, digits, and hyphens with no spaces are single tokens (e.g., *COVID-19*, *Facebook*).
6. **Numerals, measures, dates**:

  - Numeric strings with separators are single tokens (*1,680*, *2025-08-22*).
  - Units/classifiers are separate tokens: *3* ຄົນ, *1680* ກີບ.
  - Spelled-out numerals follow normal word segmentation.

7. **Named entities**: segment by word boundaries. Latin names are single tokens per contiguous sequence (e.g., *John King*).
8. **Punctuation**: each mark is its own token. Exception: decimal points within numbers (e.g., *3.14*).
9. **URLs, emails, hashtags, @mentions**: contiguous sequences are single tokens.
10. **Normalization**: text is Unicode UTF-8 NFC; zero-width/control characters are removed.

Priority and tie-breaking Apply rules in this order: (1) normalization → (2) numeric/date/url rules → (3) mixed-script cohesion → (4) function-word separation → (5) compound test → (6) punctuation separation. If multiple rules apply simultaneously, linguistic criteria take precedence over surface form. For example, ຈະພັດທະນາ is split into ຈະ (PRA) + ພັດທະນາ (V), while lexicalized compounds such as ນຳໃຊ້ remain single tokens. In borderline cases (e.g., reduplicated vs. derived forms such as ດີ ໃຈ), the final decision follows native-linguist adjudication.

Edge-case examples

- **Function-particle chain:** ໄດ້ ກໍ່ ⇒ two tokens.
- **Hybrid Lao–Latin form:** ຟຣີ-wifi ⇒ split at hyphen.
- **Borrowed numeral + classifier:** *4G* ຮ້ານ ⇒ separate tokens.
- **Nominalized verb:** ການ ຮຽນ ⇒ both tokens kept separate for syntactic clarity.

Adjudication and consistency When annotators disagreed on boundary placement, each case was reviewed by a native linguist and resolved according to written rules. All adjudication outcomes and examples of complex morphology were documented in a public changelog to ensure full transparency and reproducibility.

Appendix C. POS tagset (supplementary layer)

The POS layer serves as an auxiliary linguistic annotation added to the segmented corpus to enhance its usability for downstream NLP tasks. Each segmented token receives a single POS tag from a compact Lao-oriented tagset adapted from Universal Dependencies (UD) conventions.

All POS annotations were verified by native linguists and released publicly with the dataset (`https://github.com/HaHVU/HVULao_NLP`). The full set of POS tags and their descriptions are summarized in Table 7 below.
This supplementary POS layer is not the main contribution of the paper but strengthens the linguistic completeness and practical reusability of the HVULao_NLP corpus.

Table 7: POS tagset used in HVULao_NLP

Tag	**Description**
N	Noun
V	Verb
ADJ	Adjective
ADV	Adverb
PRE	Preposition
COJ	Conjunction
REL	Relative marker
NUM	Numeral
PUNCT	Punctuation
PRS	Pronoun
DMN	Demonstrative
NEG	Negation particle
CLF	Classifier / collective determiner
FIX	Fixed functional morpheme / nominalizer
IAC	Interjection / discourse particle
NTR	Neutral / interrogative particle (*yes/no*, *wh*-like)
PRA	Auxiliary / modal verb
PVA	Passive / causative marker
TTL	Title / honorific
IAQ	Interrogative adjunct
IBQ	Approximation / quantifier
PRN	Pronominal / referential noun (category label)
DAN	Determiner / adnominal modifier
DBQ	Distributive quantifier (e.g., reduplicated “every”)

References

[1] Anand Kumar M., “NITK-IT_NLP@NSURL2019: Transfer Learning based POS Tagger for Under Resourced Bhojpuri and Magahi Language,” in: *Proceedings of the 1st International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 – Short Papers*, Association for Computational Linguistics, Trento, Italy, 2019, pp. 68–72. `https://aclanthology.org/2019.nsurl-1.10/`

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proc. ACL, 2020, pp. 8440–8451. `https://doi.org/10.18653/v1/2020.acl-main.747`

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proc. NAACL-HLT 2019, Minneapolis, Minnesota, June 2019, pp. 4171–4186. https://aclanthology.org/N19-1423/

[4] R. Eskander, C. Lowry, S. Khandagale, J. Klavans, M. Polinsky, S. Muresan, Unsupervised Stem-based Cross-lingual Part-of-Speech Tagging for Morphologically Rich Low-Resource Languages, in: Proc. NAACL-HLT 2022, pp. 4061–4072. https://doi.org/10.18653/v1/2022.naacl-main.298

[5] R. Eskander, S. Muresan, M. Collins, Unsupervised Cross-Lingual Part-of-Speech Tagging for Truly Low-Resource Scenarios, in: Proc. EMNLP 2020, pp. 4820–4831. https://doi.org/10.18653/v1/2020.emnlp-main.391

[6] A. Imani, S. Severini, M. J. Sabet, F. Yvon, H. Schütze, Graph-Based Multilingual Label Propagation for Low-Resource Part-of-Speech Tagging, in: Proc. EMNLP 2022, pp. 1577–1589. https://doi.org/10.18653/v1/2022.emnlp-main.102

[7] T. Kudo, CRF++: Yet Another CRF Toolkit, Technical Report, 2005. http://crfpp.sourceforge.net/

[8] S. Kumar, P. Jyothi, P. Bhattacharyya, Part-of-speech Tagging for Extremely Low-resource Indian Languages, in: Findings of ACL 2024, Bangkok, Thailand, 2024, pp. 14422–14431. https://aclanthology.org/2024.findings-acl.857/

[9] C. D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, May 1999. https://icog-labs.com/wp-content/uploads/2014/07/Christopher_D._Manning_Hinrich_Sch%C3%BCtze_Foundations_Of_Statistical_Natural_Language_Processing.pdf

[10] S. Moeller, L. Liu, M. Hulden, To POS Tag or Not to POS Tag: The Impact of POS Tags on Morphological Learning in Low-Resource Settings, in: Proc. ACL-IJCNLP 2021 (Long Papers), 2021, pp. 966–978. https://doi.org/10.18653/v1/2021.acl-long.78

[11] T. Pires, E. Schlinger, D. Garrette, How Multilingual is Multilingual BERT?, in: Proc. ACL 2019, pp. 4996–5001. https://doi.org/10.18653/v1/P19-1493

[12] Y. Shen, J. Li, S. Huang, Y. Zhou, X. Xie, Q. Zhao, Data Augmentation for Low-resource Word Segmentation and POS Tagging of Ancient Chinese Texts, in: Proc. 2nd Workshop on Language Technologies for Historical and Ancient Languages, Marseille, France, 2022, pp. 169–173. https://aclanthology.org/2022.lt4hala-1.26

[13] B. Plank, A. Søgaard, Y. Goldberg, Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss, in: Proc. ACL 2016 (Short Papers), pp. 412–418. https://doi.org/10.18653/v1/P16-2067

[14] K. Phatthiyaphaibun, T. Phon-Amnuaisuk, LaoNLP: A Toolkit for Lao Natural Language Processing, Zenodo (2022). https://doi.org/10.5281/zenodo.6833407

[15] N. Goyal, J. Gao, V. Chaudhary, P. Chen, G. Wenzek, V. Ju, S. Krishnan, M. Ranzato, F. Guzmán, A. Fan, The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation, Trans. Assoc. Comput. Linguist. 10 (2022), 522–538. https://doi.org/10.1162/tacl_a_00474

[16] K. Khankasikam, N. Muansuwqan, Thai word segmentation: a lexical semantic approach, in: Proc. Machine Translation Summit X: Posters, Phuket, Thailand (2005), 331–338. https://aclanthology.org/2005.mtsummit-posters.2/

[17] N. Ljubešić, T. Erjavec, Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene, in: Proc. LREC 2016, Portorož, Slovenia, 2016, pp. 1527–1531. https://aclanthology.org/L16-1242/

[18] T. Ruokolainen, O. Kohonen, K. Sirts, S.-A. Grönroos, M. Kurimo, S. Virpioja, A Comparative Study of Minimally Supervised Morphological Segmentation, *Computational Linguistics* 42(1) (2016), 91–120. https://doi.org/10.1162/COLI_a_00243

[19] H. P. Lê, T. M. Huyền, A. Roussanaly, T. V. Hồ, A Hybrid Approach to Word Segmentation of Vietnamese Texts, in: Proc. LATA 2008, Tarragona, Spain, pp. 240–249. https://hal.inria.fr/inria-00334761

[20] X. Pan, B. Zhang, Y. Wang, J. Chen, Y. Shen, Cross-lingual Transfer for Low-resource Asian Languages: Challenges and Opportunities, in: Proc. COLING 2020, pp. 2706–2717. https://aclanthology.org/2020.coling-main.243

[21] C. Haruechaiyasak, C. Kongthon, A. Sangkeettrakarn, A. Palingoon, A Comparative Study on Thai Word Segmentation Approaches, in: Proc. IJCNLP 2008, pp. 282–287. https://www.cs.ait.ac.th/~mdailey/papers/Choochart-Wordseg.pdf

[22] N. Chanta, V. Theeramunkong, Thai Word Segmentation based on Global and Local Unsupervised Learning, *Informatica* 30(4) (2006), 403–414.

[23] Y. Zhou, Y. Zhang, P. Li, Optimal word segmentation for neural machine translation into Dravidian languages, *Proc. 8th Workshop on Asian Translation (WAT 2021)*, 205–214 (2021). `https://aclanthology.org/2021.wat-1.21/`

[24] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The State and Fate of Linguistic Diversity and Inclusion in the NLP World, in: Proc. ACL 2020, pp. 6282–6293. `https://doi.org/10.18653/v1/2020.acl-main.560`

[25] H. Kaing, C. Ding, M. Utiyama, E. Sumita, S. Sam, S. Seng, K. Sudoh, S. Nakamura, Towards Tokenization and Part-of-Speech Tagging for Khmer: Data and Discussion, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 20 (6) (2021) Article 104. `https://doi.org/10.1145/3464378`

[26] C. Ding, Y.K. Thu, Word Segmentation for Burmese (Myanmar), ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2016). `https://doi.org/10.1145/2846095`

[27] W.P. Pa, H.A. Oo, T. San, Word Boundary Identification for Myanmar Text Using Conditional Random Fields, in: Proc. Int. Conf. Asian Lang. Process. (IALP), IEEE, 2015. `https://doi.org/10.1007/978-3-319-23207-2_46`

[28] C. Mao, Z. Man, Z. Yu, H. Wang, A Neural Joint Model with BERT for Burmese Syllable Segmentation, Word Segmentation, and POS Tagging, ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2021). `https://doi.org/10.1145/3436818`

[29] Y. Li, X. Li, Y. Wang, H. Lv, F. Li, L. Duo, Character-based Joint Word Segmentation and Part-of-Speech Tagging for Tibetan Based on Deep Learning, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21 (5) (2022) Article 95. `https://doi.org/10.1145/3511600`

[30] Z. Wang, Y. Lyu, J. Zhu, Tibetan Word Segmentation Based on Word-Position Tagging, in: Proc. Int. Conf. Asian Lang. Process. (IALP), IEEE, 2013. `https://doi.org/10.1109/IALP.2013.74`

[31] R. Buoy, N. Taing, S. Kor, Joint Khmer Word Segmentation and Part-of-Speech Tagging Using Deep Learning, arXiv preprint (2021). `https://arxiv.org/abs/2103.16801`

[32] J. Hu, J. Fu, W. Zhao, P. Lou, M. Feng, H. Ren, A. Fang, Characterizing pituitary adenomas in clinical notes: Corpus construction and its application in LLMs, *Health Informatics Journal* 30(4) (2024), 14604582241291442. `https://doi.org/10.1177/14604582241291442`

[33] A. Ahmad, M. Azzeh, E. Alnagi, Q. Abu Al-Haija, D. Halabi, A. Aref, Y. AbuHour, Hate speech detection in the Arabic language: corpus design, construction, and evaluation, *Frontiers in Artificial Intelligence* 7 (2024), 1345445. `https://doi.org/10.3389/frai.2024.1345445`

[34] R. Shao, P. Lin, Z. Xu, Integrated natural language processing method for text mining and visualization of underground engineering text reports, *Automation in Construction* 166 (2024), 105636. `https://doi.org/10.1016/j.autcon.2024.105636`