When Morphology Hides in Plain Sight: Breaking the Isolation in Vietnamese and Beyond

Anh Trac Duc Dinh[1], Khang Hoang Nhat Vo[2], Tai Tien Ta[1]
Vinh Cong Doan[1], Tho Quan[1*]

[1]Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam
[2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[*]**Corresponding author:** qttho@hcmut.edu.vn

Abstract

In isolating languages such as Vietnamese, core morphological structure is encoded not by inflection but by the composition and ordering of monosyllabic morphemes, yet standard Transformer encoders largely overlook this signal. We introduce **HuTieuBERT**[1], a morpheme-aware Transformer that augments a pretrained Vietnamese encoder with two lightweight inductive biases: (i) *Adaptive Boundary-Token Fusion*, which integrates BMES-based morpheme boundary embeddings into token representations via a learnable gate, and (ii) a *Morpheme-Aware Attention Bias*, which injects a fixed structural attention matrix into early self-attention layers while minimally perturbing the pretrained attention geometry. Across a suite of Vietnamese POS, NER, and sentence-level classification benchmarks, HuTieuBERT consistently outperforms strong baselines, with the largest gains on syntactic tasks. Hyperparameter ablations show a broad regime in which structural biases improve accuracy without destabilizing representations. Applying the same design to ChineseBERT (ChineseBERT-wwm) yields **MAChineseBERT**, which improves F_1 and produces more balanced tag distributions on Chinese POS and NER, suggesting that explicit morpheme-aware attention is a portable and effective strategy for modeling isolating languages.

1 Introduction

Vietnamese belongs to the family of isolating languages (Do, 2013; Do-Hunville and Dao, 2018), where words remain in base form without inflection and meaning is constructed by combining independent morphemes (Than, 1997; Le, 2003; Chau, 2007). Each morpheme is typically monosyllabic and written as a sequence of continuous characters (Bloomfield, 1933). In sentences, morphemes are arranged left to right and separated by spaces.

This linguistic profile aligns with the Isolating Monocategorial Associational language typology (Gil, 2006), characterized by three core properties that influence all levels of Vietnamese language processing, with most direct impact on syntactic tasks where morpheme boundaries determine structural annotations. First, *morphological isolation* is evident in the invariant form of words across grammatical contexts (e.g., *đi* "go" remains *đi* regardless of tense or person). Second, *syntactic monocategoriality* enables *transcategoriality* (Do-Hunville and Dao, 2018), where morphemes function flexibly across word classes depending on context (e.g., *ăn* may denote to "eat" as a verb, refer to *thức ăn* "food" as a noun, or appear in compounds like *ăn chay* "vegetarian diet"). Third, *semantic associationality* makes morpheme order consequential for lexical semantics (Hiep, 2020), as reordering produces different word meanings (e.g., *học sinh* "student" vs. *sinh học* "biology", *quân đội* "army" vs. *đội quân* "troop/unit", *nhà nước* "the state" vs. *nước nhà* "the homeland", and *quốc dân* "citizens" vs. *dân quốc* "republic").

Morphemes form fixed compounds (e.g., *xe máy* "motorbike") and convey grammatical information through function words (e.g., *đã đi* "(have) gone"). Multi-syllabic expressions often exhibit boundary ambiguity (Dinh et al., 2008), functioning as single lexical units despite each syllable carrying independent meaning (e.g., *máy tính xách tay* "laptop", literally "computer carry-hand"). Consequently, tasks such as POS tagging, NER, and parsing rely heavily on correctly identifying morphemic boundaries and their interactions (Bach et al., 2013; Nguyen et al., 2019).

Current Transformer Encoder-based Vietnamese models do not explicitly capture morphemic relations. Even segmentation-based models like PhoBERT (Nguyen and Tuan Nguyen, 2020) and BARTpho (Tran et al., 2022) group syllables into tokens without adapting self-attention to mor-

[1]The code is available at https://github.com/TracDucAnh/HuTieuBERT.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10377–10392

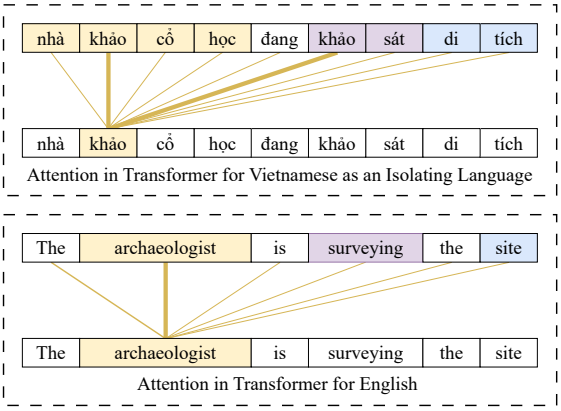

Figure 1: Attention diffusion in Vietnamese vs. English. In *nhà khảo cổ học đang khảo sát di tích* ("the archaeologist is surveying the site"), Vietnamese attention spreads across morpheme boundaries (*khảo* in *nhà khảo cổ học* "archaeologist" attends to *khảo* in *khảo sát* "survey"), whereas English attention remains localized along syntactic dependencies.

phemic structure. As a result, attention may spread across morpheme boundaries without respecting compound structure, reducing effectiveness for structure-sensitive tasks (Figure 1).

We introduce HuTieuBERT, a morpheme-aware Transformer Encoder[2] that augments pretrained encoders with explicit morphological structure by encoding morpheme boundaries and compound cohesion directly in the architecture (Figure 2). This approach produces representations better aligned with Vietnamese and typologically related isolating languages such as Mandarin Chinese (Erbaugh, 2022; Zhang, 2023) and Thai (Sornlertlamvanich and Pantachat, 1993).

2 Literature Review

2.1 The Morphemic Foundations of Vietnamese

Vietnamese is widely characterized as a morpheme-based language in which the fundamental semantic unit is the morpheme, typically realized as a monosyllable (Than, 1997; Chau, 2007; Can, 1996). Each morpheme carries lexical or grammatical meaning, and any further segmentation yields only phonemes without semantic content. Consequently, the interpretation of a word depends on how its constituent morphemes are combined, ordered, and delimited.

[2]The model name is inspired by *hủ tiếu*, a popular Southern Vietnamese noodle dish originating from Teochew (Chaozhou), China.

The literature identifies three central morphological mechanisms in Vietnamese.

Morpheme lexicalization. Many morphemes function as independent words, such as *ăn* ("to eat") and *đi* ("to go"), whose meanings are complete at the monosyllabic level.

Compounding. Morphemes combine to form new lexical units, e.g., *học sinh* ("student", from *học* "study" + *sinh* "person") and *xe máy* ("motorbike", from *xe* "vehicle" + *máy* "engine"). Two main types exist: *coordinating compounds*, where morphemes contribute equally (e.g., *núi non* "mountains"), and *subordinating compounds*, where a head is refined by a dependent (e.g., *nhà bếp* "kitchen", with *nhà* "house" as head and *bếp* "cooking" as modifier).

Reduplication. Reduplication creates expressive or descriptive nuances, as seen in *luôn luôn* ("always") through full repetition and *lấp lánh* ("glittering") through partial repetition (Can, 1996).

These features highlight morphemes' central role in Vietnamese lexical semantics and motivate morpheme-aware modeling. However, some scholars advocate a word-based view (Hao, 1998; Giap, 2008, 2011; Hieu Nguyen et al., 2025), questioning morphemes' semantic autonomy; importantly, (Hieu Nguyen et al., 2025) challenges the assumption that words are built from syllables, showing the concept remains debated.

We take a morpheme-based approach because it fits Vietnamese's monosyllabic, semantically rich structure and is well supported by empirical evidence, providing a solid basis for models that incorporate morphemic information.

2.2 The role of segmentation in Vietnamese syntactic analysis

Word segmentation is essential in Vietnamese NLP, an isolating language where syllables surface individually and multi-syllabic words form via morpheme compounding (Dien et al., 2001; Nguyen and Le, 2016). Proper segmentation preserves lexical boundaries critical for syntax.

Early supervised approaches established foundational methods. Nguyen et al. (2006) investigated CRFs and SVMs for Vietnamese word segmentation, showing that word boundaries depend more on syllable identity than positional features, with SVM-based learning proving particularly effective. Vu et al. (2018) advanced implementation with

VnCoreNLP's RDRSegmenter, achieving state-of-the-art performance using Ripple Down Rules on the Vietnamese Treebank (Nguyen et al., 2009).

Critically, Nguyen and Tuan Nguyen (2020) revealed that existing BERT (Devlin et al., 2019) models failed to capture syllable-word distinctions. PhoBERT addressed this by requiring segmentation preprocessing to encode multi-syllabic units, while BARTpho extended this to sequence-to-sequence tasks. These studies established explicit segmentation as essential for Vietnamese pre-trained models.

Recent work challenges this paradigm. Hieu Nguyen et al. (2025) propose ViWordFormer, a segmentation-free Transformer with phrasal-lexeme and co-text modules that models word formation from syllables. They find segmentation-free models match or surpass segmented baselines on NLU tasks. However, comparisons were limited to simple architectures (Standard Transformer, LSTM, GRU, TextCNN), leaving unclear whether segmentation benefits persist with modern pre-trained models like PhoBERT or BARTpho.

2.3 Attention-based Methods for Vietnamese Morphological Structure

Attention mechanisms have revolutionized NLP by enabling models to capture long-range semantic dependencies (Vaswani et al., 2017; Devlin et al., 2019). However, their application to isolating languages like Vietnamese presents unique challenges. Unlike inflecting languages where morphological markers explicitly signal grammatical relationships, Vietnamese relies on word order and multi-syllable compounds without inflectional morphology (Suni, 2007; Phan and Lander, 2015; Bui, 2019). This requires attention mechanisms to precisely focus on relevant morpheme boundaries within compounds.

Analyses show that many BERT attention heads have diffuse patterns: some allocate up to 50% of attention to [SEP] tokens, while others distribute attention nearly uniformly. Only certain heads in specific layers (e.g., layer 8, heads $0, 1$) reliably focus on syntactic tokens (Clark et al., 2019). This behavior likely extends to BERT-derived models like RoBERTa (Liu et al., 2019) and PhoBERT, which lack explicit structural biases. In isolating languages such as Vietnamese, diffuse attention can weaken focus on morphologically coherent units (Chen and Allport, 1995; Gil, 2008), motivating the use of structurally-guided attention.

Recent work on Vietnamese NLP employs various attention mechanisms: label attention with CRF (85% F1 on parsing (Vu-Tran et al., 2022)), self-attention Transformers (91% on sentiment (Nguyen et al., 2020)), and word-level attention (relation extraction (Nguyen et al., 2018b)). However, these face key limitations. While Vietnamese-specific models like PhoBERT and BARTpho treat segmented words (e.g., *cà_phê* "coffee") as atomic tokens concatenated by "_" and multilingual models (mBERT (Pires et al., 2019), XLM-RoBERTa (Conneau et al., 2020)) lack isolating-language optimization, all lack explicit structural biases to encode morpheme boundaries within compounds. Chinese BERT-wwm (Cui et al., 2020) extends this approach by treating multi-character words as cohesive units through Whole Word Masking, yet similarly does not incorporate explicit morphological structure into attention, motivating our morpheme-aware attention approach.

2.4 Cross-linguistic Morphological Properties

While our work centers on Vietnamese, the core morphological challenges extend to many isolating languages. These languages share key characteristics (Luo, 2020): monosyllabic morphemes, compounding as the primary word-formation mechanism, and grammatical encoding through word order and function words rather than inflection, tendencies well-established in linguistic typology.

Mandarin Chinese is a prototypical example of this profile. Like Vietnamese, it lacks tense inflection and instead uses aspectual markers (Lin, 2006) - for instance, *le* (了) for completed events and *zài* (在) for ongoing actions, paralleling Vietnamese *đã* and *đang*. Both languages also exhibit *transcategoriality*, where a single lexical item can function across parts of speech. For example, a Chinese word 学习 can appear as a verb "to study" or as a noun "learning", analogous to Vietnamese *học*. As a verb: 我学习汉语 ("I study Chinese"). As a noun: 学习很重要 ("Learning is important").

Word order sensitivity also creates challenges. In Chinese 故事 ("story") vs. 事故 ("accident") illustrate how morpheme order alone can induce substantial semantic shifts (Xiong et al., 2015; Zhou and Liu, 2022).

These shared properties imply that Vietnamese-focused computational methods, such as morpheme-aware attention, can generalize to Chinese and other isolating languages.

3 Proposed Methodology

3.1 Adaptive Boundary-Token Fusion

3.1.1 Embedding Module

We segment sentences into words using language-specific segmenters (VnCoreNLP for Vietnamese, Jieba[3] for Chinese) and produce BMES tags for each syllable. For subword tokenization, we align BMES tags to subwords before feeding them with BMES embeddings into the Transformer Encoder (Figure 2).

The Adaptive Boundary-Token Fusion module combines a boundary embedding E_b with a token embedding E_t to obtain a fused representation E_f via a learnable, element-wise gating vector W_i:

$$W_i = \sigma\Big(W_\lambda[E_b; E_t] + b_\lambda\Big), \tag{1}$$

$$E_f = W_i \odot E_b + (1 - W_i) \odot E_t, \tag{2}$$

$$E_{\text{LN}} = \text{LayerNorm}(E_f), \tag{3}$$

where E_b and E_t denote the boundary and token embeddings, respectively, $[E_b; E_t]$ is their concatenation, and $\sigma(\cdot)$ is the sigmoid function. The parameters W_λ and b_λ are learned during training and determine the interpolation weights W_i. The vector E_f is the fused embedding, and E_{LN} is the final representation after layer normalization.

3.1.2 BMES-Subword Alignment

Subword tokenizers may split words into multiple subwords, misaligning with word-level BMES tags. We address this by expanding each tag to subword level. For a word tokenized into n subwords, the subword tag y'_i is:

$$y'_i = \begin{cases} B, & y_i = B \wedge i = 1, \\ E, & y_i = E \wedge i = n, \\ M, & y_i \in \{B, E\} \wedge 1 < i < n, \\ y_i, & y_i \in \{S, M\}, \end{cases} \tag{4}$$

where y_i is the original word-level tag and n is the number of subwords. For example, *Lập trình viên* ("programmer") with original B-M-E labels tokenizes to [L@@, ập@@, trình, vi@@, ên@@] and maps to B-M-M-M-E, preserving morphemic structure at subword level.

[3] https://github.com/fxsjy/jieba

3.2 Morpheme-Aware Attention Bias

3.2.1 Structural Attention Matrix

To encourage self-attention to respect morphemic and phrasal boundaries, we introduce a fixed *Structural Attention Matrix* $\mathbf{B} \in \mathbb{R}^{T\times T}$, where each entry $\mathbf{B}_{i,j}$ modulates the attention score between tokens i and j according to their structural relationship:

$$\mathbf{B}_{i,j} = \alpha\,\mathbf{1}_{i\sim j} + \beta\,\mathbf{1}_{i\not\sim j} + \gamma\,\mathbf{1}_{i,j\in S} + \delta\,\mathbf{1}_{i=j}, \tag{5}$$

where the indicator functions are defined as follows: $\mathbf{1}_{i\sim j} = 1$ if tokens i and j belong to the same compound phrase and 0 otherwise; $\mathbf{1}_{i\not\sim j} = 1$ if tokens i and j belong to different compounds and 0 otherwise; $\mathbf{1}_{i,j\in S} = 1$ if both tokens are single-word units (S) and 0 otherwise; and $\mathbf{1}_{i=j} = 1$ if $i = j$ (self-attention bias) and 0 otherwise (See Appendix D for illustrative examples).

We constrain $\alpha \geq 0$, $\beta \leq 0$, $\gamma < \alpha$, and use δ to control self-attention strength. Intuitively, α strengthens within-compound attention, β penalizes cross-compound attention, γ highlights single-word tokens, and δ adjusts self-attention bias. These parameters are fixed to provide a stable, interpretable structural prior.

3.2.2 Application to Attention Heads

We apply the Structural Attention Matrix $\mathbf{B}$ to a subset of attention heads. Let $\mathcal{S}(\cdot)$ denote the softmax operation (Figure 2). The biased multi-head attention for head h is defined as:

$$\mathcal{A}_h = \mathcal{S}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}} + \delta_{h,h^*}\,\mathbf{B}\right)V_h, \tag{6}$$

where Q_h, K_h, and V_h are the query, key, and value matrices for head h, respectively, and d_k is the key dimensionality. The Kronecker delta δ_{h,h^*} ensures that only the designated structural head h^* incorporates the bias matrix $\mathbf{B}$. Here, $\mathcal{A}_h$ denotes the output of head h. The matrix $\mathbf{B}$ is constructed once per input sequence from BMES tags and reused across all layers.

3.2.3 Choosing Layers to Inject Structural Attention Matrix

Probing studies indicate that BERT layers gradually encode linguistic information, from surface and morphological patterns in early layers to syntax and semantics in later layers (Jawahar et al., 2019; Lin et al., 2019), though the hierarchy is not

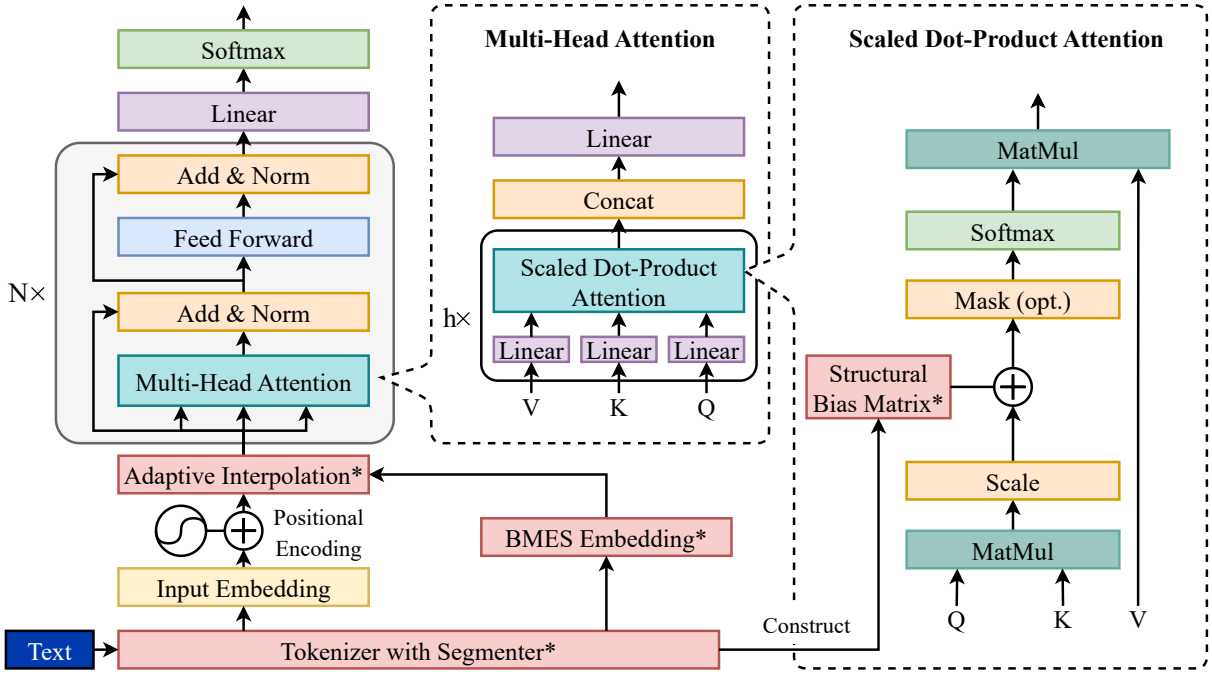

Figure 2: Morpheme-Aware Transformer Encoder. Standard Transformer encoder augmented with morpheme boundary information. Novel components (*) include Tokenizer with Segmenter, BMES Embedding, Adaptive Interpolation (Adaptive Boundary-Token Fusion), and Structural Bias Matrix (Structural Attention Matrix). Other components follow standard Transformer design.

perfectly separated (Nikolaev and Padó, 2023; Niu et al., 2022). Morphological–syntactic cues are thus concentrated in early–middle layers.

Motivated by these findings, we partition BERT-base model (Devlin et al., 2019) into three stages: *early* (layers $0-3$), *middle* ($4-7$), and *deep* ($8-11$). We inject the Structural Attention Matrix into a subset of inner layers ($1-2$, $5-6$, $9-10$), while leaving the boundary layers of each stage unchanged to better preserve the pretrained representations. Empirically, injecting structural bias into the lowest layers ($1-2$, applied to all 12 heads) yields the largest improvements on Vietnamese syntactic benchmarks, as reported in our experimental results and later confirmed by our ablation study. This pattern aligns with prior evidence that early Transformer layers are particularly sensitive to morphological and local syntactic information.

4 Experiments

4.1 Baselines

We compare HuTieuBERT against widely used pretrained BERT-based encoders, including PhoBERT-base, XLM-RoBERTa-base, and mBERT. Although ViWordFormer (Hieu Nguyen et al., 2025) is closely related in spirit, we do not include it in our main comparison due to substantial architectural and pretraining differences: ViWordFormer is a segmentation-free model trained from scratch, whereas HuTieuBERT augments pretrained encoders, making direct and controlled evaluation less informative.

4.2 Dataset

Intermediate task. To expose the model to broad Vietnamese usage, we further pre-train the encoder with masked language modeling (MLM) on the 2GB Vietnamese Book Corpus[4] (∼4M sentences, mean length 256 tokens), enabling HuTieuBERT to acquire general linguistic regularities beyond the original pretrained encoder.

Syntactic tasks. We evaluate two sequence-labeling tasks: POS and NER. For POS, we use UD Vietnamese TreeBank (UD_VTB), VnDT (Nguyen et al., 2014), and the VLSP 2013 (VLSP_13) dataset[5]. For NER, we employ PhoNER_COVID19 (PhoNER) (Truong et al., 2021), VietMed-NER (VietMed) (Le-Duc et al., 2025), and the VLSP

[4] https://huggingface.co/datasets/tmnam20/Vietnamese-Book-Corpus

[5] https://vlsp.org.vn/vlsp2013/eval/ws-pos

Table 1: Performance on Vietnamese NLP Tasks.

Task	Dataset	HuTieuBERT	PhoBERT	XLM-RoBERTa	mBERT
Syntactic Tasks (Acc. for POS and F1 for NER)					
POS	UD_VTB	**0.9567** $\pm$ 0.0031	0.9188 $\pm$ 0.0015	0.9189 $\pm$ 0.0165	0.9049 $\pm$ 0.0007
	VnDT	**0.9873** $\pm$ 0.0006	0.9465 $\pm$ 0.0008	0.8936 $\pm$ 0.0070	0.9084 $\pm$ 0.0028
	VLSP_13	**0.9654** $\pm$ 0.0003	0.9601 $\pm$ 0.0009	0.9577 $\pm$ 0.0008	0.9607 $\pm$ 0.0030
NER	PhoNER	**0.8855** $\pm$ 0.0011	0.8383 $\pm$ 0.0015	0.6975 $\pm$ 0.0038	0.7282 $\pm$ 0.0043
	VietMed	0.5891 $\pm$ 0.0089	0.5604 $\pm$ 0.0444	**0.6176** $\pm$ 0.0088	0.5981 $\pm$ 0.0057
	VLSP_16	**0.9235** $\pm$ 0.0041	0.9227 $\pm$ 0.0050	0.7564 $\pm$ 0.0049	0.7685 $\pm$ 0.0057
Semantic Tasks (F1)					
Sent. Ana.	VFSC	**0.8317** $\pm$ 0.0044	0.8095 $\pm$ 0.0126	0.8057 $\pm$ 0.0172	0.7830 $\pm$ 0.0146
Topic Cls.	VFSC	**0.8042** $\pm$ 0.0125	0.7878 $\pm$ 0.0071	0.7766 $\pm$ 0.0179	0.7580 $\pm$ 0.0071
Cons. Det.	ViCTSD	0.8141 $\pm$ 0.0086	**0.8211** $\pm$ 0.0100	0.8162 $\pm$ 0.0061	0.8162 $\pm$ 0.0072
Toxic Det.	ViCTSD	**0.7415** $\pm$ 0.0082	0.7253 $\pm$ 0.0166	0.7329 $\pm$ 0.0209	0.6853 $\pm$ 0.0367

2016 (VLSP_16) dataset[6]. We follow the original data splits for all datasets except VLSP_16, for which no official development set is provided; thus, we randomly sample 15% of the training set as development data using a fixed seed of 42.

Semantic tasks. We consider sentence-level classification benchmarks. UIT-VFSC (VFSC) (Nguyen et al., 2018a) is used for Sentiment Analysis (Sent. Ana.) and Topic Classification (Topic Clas.), while UIT-ViCTSD (ViCTSD) (Nguyen et al., 2021) is used for Constructive Detection (Cons. Det.) and Toxic Detection (Toxic Det.).

4.3 Two-Stage Fine-tuning and Training Configurations

All models are optimized using a two-stage fine-tuning procedure. Only datasets at the sentence level are word-segmented using VnCoreNLP (required for PhoBERT/HuTieuBERT); datasets that are already segmented require no preprocessing.

In stage 1, we continue MLM training on the Vietnamese Book Corpus for 5 epochs, initializing from PhoBERT with the encoder frozen for the first epoch then unfrozen. In stage 2, we fine-tune for 30 epochs (syntactic tasks) or 5 epochs (semantic tasks) using AdamW with linear warmup and early stopping. We report mean $\pm$ standard deviation across five independent trials with shuffled training data.

[6] https://vlsp.org.vn/vlsp2016/eval/ner

For HuTieuBERT, we set $\alpha = 0.5$, $\beta = -0.3$, and $\gamma = \delta = 0$ based on ablation studies (Section 5.2). This yields cosine similarity $C \approx 0.9995$ and Frobenius norm $\|\Delta\|_F \approx 0.1447$ between biased and unbiased attention matrices, indicating strong structural guidance while preserving pre-trained geometry. We apply the bias to layers $1 - 2$ across all 12 heads.

4.4 Results

Table 1 summarizes the performance of HuTieuBERT across all Vietnamese benchmarks, highlighting particularly strong gains on syntactic tasks.

Syntactic tasks. HuTieuBERT achieves the best performance among all evaluated models on POS tagging, with accuracies of 95.67% on UD_VTB, 98.75% on VnDT, and 96.54% on VLSP 2013. These results correspond to absolute improvements over PhoBERT of 3.79, 4.10, and 0.53 percentage points, respectively, and are accompanied by very low variance (standard deviations in the range $0.0006 - 0.0031$), indicating highly stable training. For NER, HuTieuBERT attains 88.55% on PhoNER and 92.35% on VLSP 2016, consistently outperforming all baselines. On VietMed, HuTieuBERT reaches 58.91%, lagging behind XLM-RoBERTa (61.76%). We attribute this gap to the dataset's heavy use of domain-specific English medical terminology (e.g., l-carnitine fumarate, betacarotene, phytoncide) and pronounced class imbalance (78.48% O-tokens), conditions under which multilingual models can better exploit cross-

lingual lexical knowledge.

Semantic tasks. On sentence-level classification, HuTieuBERT obtains 83.17% for Sentiment Analysis and 80.42% for Topic Classification, outperforming PhoBERT by 2.22 and 1.64 percentage points, respectively. These gains suggest that explicitly modeling multi-syllabic compounds (e.g., *phấn khích* "exciting", *công nghệ* "technology") is beneficial even for sentence-level semantics. For toxicity-related benchmarks, HuTieuBERT achieves the highest performance on Toxic Detection (74.15%), while its score on Constructiveness Detection (81.41%) is comparable to PhoBERT (82.11%). The latter task appears to rely more heavily on broader discourse-level context than on local phrasal patterns, which may limit the impact of morpheme-aware modeling.

5 Ablation Studies

5.1 Architecture-Level Ablation

We evaluate structural bias injection at three depths (layers $1-2$, $5-6$, and $9-10$), applying it to all 12 attention heads in the selected layers. To isolate each component's contribution, *w_bias_1-2* uses only the Structural Attention Matrix on layers $1-2$, while *w_bmes* uses only Adaptive Boundary–Token Fusion with BMES embeddings.

Table 2 shows that early-layer injection $(1-2)$ generally outperforms deeper injection on POS tasks. On UD_VTB, layers $1-2$ improve accuracy by $0.3-0.6\%$ over deeper variants, though the advantage is marginal on VnDT ($< 0.1\%$). For NER, layer selection has less impact on PhoNER and VLSP_16 (within $0.2-0.9\%$), but shows larger differences on VietMed (up to 2.0%).

Removing either component degrades performance. The *w_bias_1-2* variant (without BMES fusion) drops $0.4-1.1\%$ on POS tasks, while w_bmes (without structural bias) shows $0.2-0.4\%$ decreases, indicating the Adaptive Boundary–Token Fusion contributes more substantially to syntactic parsing. On NER, component importance varies by dataset: both show minimal impact on PhoNER ($< 0.1\%$), but contribute noticeably to VietMed ($1.8-2.1\%$ drops) and VLSP_16 ($0.6-1.1\%$ drops).

Overall, these findings support two conclusions: (1) injecting structural bias in early layers best trades off respecting the pretrained representations with introducing useful morphological and phrasal constraints; and (2) both the Structural Attention Matrix and the Adaptive Boundary-Token Fusion are necessary for optimal performance, with BMES fusion being particularly beneficial for Vietnamese syntactic tasks.

5.2 Hyperparameter Ablation

The Structural Attention Matrix introduces two fixed biases, α (intra-compound) and β (inter-compound), applied to layers 1-2 full heads. We sweep $\alpha \in [0, 5]$ and $\beta \in [-5, 0]$ with step 0.5 (with $\gamma = \delta = 0$) and compare the original attention A_{orig} with the biased attention A_{bias} on $10{,}000$ Vietnamese Book Corpus sentences using cosine similarity $C = \cos(A_{\text{orig}}, A_{\text{bias}})$ and Frobenius norm $\|\Delta\|_F = \|A_{\text{bias}} - A_{\text{orig}}\|_F$ (Figure 3).

As shown in Figure 3, cosine similarity remains ≥ 0.98 for a wide range of (α, β), while $\|\Delta\|_F$ grows smoothly with the magnitude of the biases. Our chosen setting $(\alpha = 0.5, \beta = -0.3)$ lies in a low-perturbation region (with $C \approx 0.9995$, $\|\Delta\|_F \approx 0.1447$) and yields the best syntactic performance (Table 1), indicating that modest structural biases are sufficient to inject useful morphological information without destabilizing the pretrained attention patterns. A more detailed analysis is provided in Appendix A.

6 Potential for Other Isolating Languages

Under the same configuration and hyperparameters described in Section 4.3, we transfer our method to ChineseBERT (ChineseBERT-wwm), yielding a morpheme-aware variant that we denote MAChineseBERT.

We adopt a two-stage continued pretraining regime with whole-word-masked MLM. For the pretraining stage, we use the THUCTC corpus (Sun et al., 2016). For evaluation, we consider two benchmarks for Simplified Chinese: UD Chinese GSDSimp (GSDS) for POS tagging (Nivre et al., 2020) and ULNER for NER (Xu et al., 2025).

Table 3 indicates that the proposed morpheme-aware architecture can be effectively transferred to Simplified Chinese. On ULNER, MAChineseBERT achieves an F1 score of 75.71%, outperforming ChineseBERT (74.40%) by 1.31 percentage points.

On GSDS, MAChineseBERT attains 81.58% accuracy, slightly below the ChineseBERT baseline (82.49%). Following standard practice, we report accuracy as the primary metric for POS tagging. However, as an auxiliary analysis to

Table 2: Performance on POS (Acc.) and NER (F1). Top 3 rows: POS tagging tasks; Bottom 3 rows: NER tasks.

	HuTieuBERT				
Dataset	**1-2**	**5-6**	**9-10**	**w_bias_1-2**	**w_bmes**
UD_VTB	**0.9567** ± 0.0031	0.9538 ± 0.0007	0.9507 ± 0.0046	0.9461 ± 0.0008	0.9530 ± 0.0005
VnDT	**0.9873** ± 0.0006	0.9868 ± 0.0009	0.9871 ± 0.0003	0.9788 ± 0.0109	0.9832 ± 0.0055
VLSP_13	**0.9654** ± 0.0003	0.9633 ± 0.0007	0.9642 ± 0.0004	0.9617 ± 0.0011	0.9634 ± 0.0005
PhoNER	0.8855 ± 0.0011	**0.8869** ± 0.0017	0.8860 ± 0.0014	0.8847 ± 0.0043	0.8860 ± 0.0014
VietMed	**0.5891** ± 0.0089	0.5691 ± 0.0052	0.5738 ± 0.0021	0.5680 ± 0.0053	0.5715 ± 0.0039
VLSP_16	**0.9235** ± 0.0041	0.9144 ± 0.0055	0.9158 ± 0.0047	0.9129 ± 0.0035	0.9176 ± 0.0024

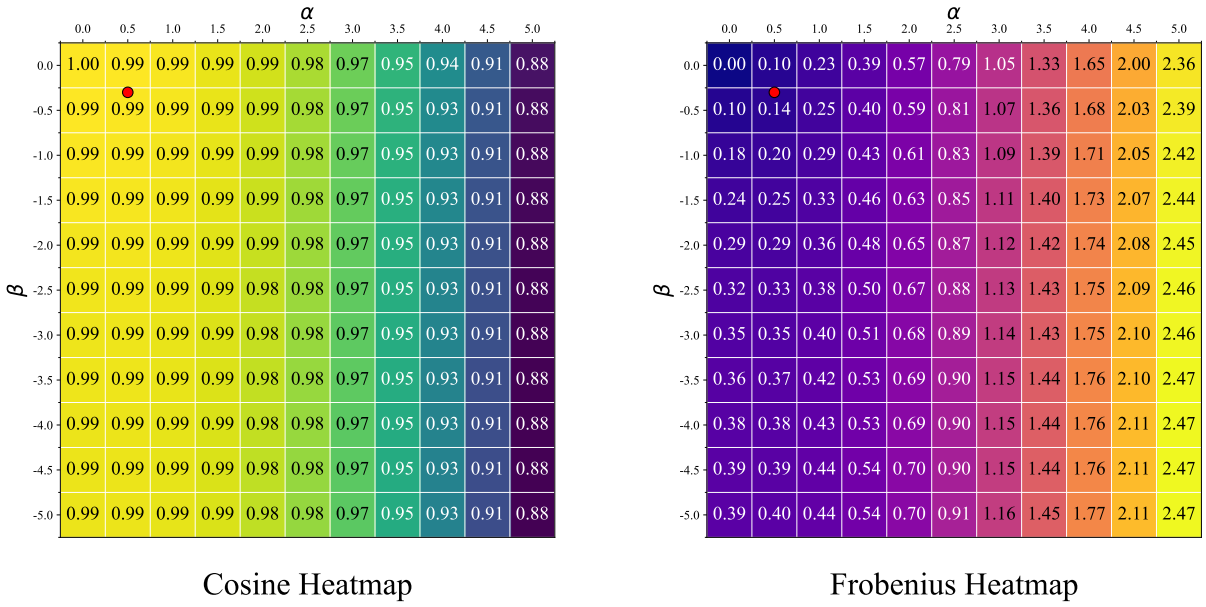

Figure 3: Cosine similarity (left) and Frobenius norm (right) over the (α, β) grid. The red dot marks the setting $(\alpha = 0.5, \beta = -0.3)$ used in our main experiments.

Table 3: Transfer to Simplified Chinese: performance on GSDS (POS) and ULNER (NER).

Dataset	**MAChineseBERT**	**ChineseBERT**
GSDS (Acc.)	0.8158 ± 0.0002	**0.8249** ± 0.0006
GSDS (F1)	**0.8316** ± 0.0005	0.7816 ± 0.0030
ULNER (F1)	**0.7571** ± 0.0030	0.7440 ± 0.0036

better understand model behavior, we also compute macro-averaged F1. Here, MAChineseBERT reaches 83.16% F1, compared to 78.16% for ChineseBERT, a gain of 5.00 percentage points.

This discrepancy between accuracy and F1 - a modest 0.91 pp decrease in accuracy alongside a substantial 5.00 pp increase in F1 - suggests that MAChineseBERT makes more balanced predictions across POS categories. In other words, the morpheme-aware attention appears to reduce the model's bias toward majority tags while improving performance on minority categories, a desirable property in the presence of label imbalance. These results provide evidence that structural biases informed by morpheme boundaries can benefit not only Vietnamese but also other isolating languages such as Simplified Chinese.

7 Conclusion

Isolating languages, which lack rich inflectional morphology and rely heavily on compounding, benefit from explicit morphological modeling. We introduce HuTieuBERT, which encodes morpheme boundaries and compound cohesion via Adaptive Boundary-Token Fusion and structural attention biases, yielding consistent improvements over strong baselines on Vietnamese syntactic tasks. When

transferred to Simplified Chinese, performance slightly decreases in accuracy but improves in F1, suggesting more balanced class-wise predictions.

8 Limitations

Our approach has several limitations that open up avenues for future research. First, the Structural Attention Matrix relies on manually chosen hyperparameters $(\alpha, \beta, \gamma, \delta)$ that remain fixed during training. We do not explore variants with learnable structural parameters, adaptive schedules, or task-specific tuning, nor do we systematically investigate the role of γ and δ beyond the setting used in our main experiments. Studying learned or dynamically modulated structural biases could reveal richer interactions between morphology and attention.

Second, our cross-lingual evaluation is restricted to Vietnamese and Simplified Chinese. While both are typologically isolating, other languages such as Thai, Lao, or certain creole languages exhibit related but distinct morphological and orthographic properties. Extending HuTieuBERT-style architectures to a broader set of isolating languages, and to multilingual pretraining scenarios, would provide a more comprehensive assessment of generalizability.

Finally, our experiments focus primarily on sequence labeling and sentence-level classification. It remains an open question how morpheme-aware attention affects higher-level tasks such as dependency parsing, natural language inference, or generative modeling. We leave the exploration of these directions, as well as potential integration with joint segmentation-tagging frameworks, to future work.

Acknowledgments

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

References

Ngo Xuan Bach, Kunihiko Hiraishi, Nguyen Le Minh, and Akira Shimazu. 2013. Dual decomposition for vietnamese part-of-speech tagging. *Procedia Computer Science*, 22:123–131.

Leonard Bloomfield. 1933. *Language*. Henry Holt, New York.

Thuy Bui. 2019. Chapter 6. temporal reference in vietnamese. *Interdisciplinary Perspectives on Vietnamese Linguistics*.

Nguyen Tai Can. 1996. *Ngữ pháp tiếng Việt (Vietnamese Grammar)*. Vietnam National University Press, Hanoi, Vietnam.

Do Huu Chau. 2007. *Từ vựng ngữ nghĩa tiếng Việt (Vietnamese Lexical Semantics)*. Vietnam Education Publishing House, Hanoi, Vietnam.

Yiping Chen and Alan Allport. 1995. Attention and lexical decomposition in chinese word recognition: Conjunctions of form and position guide selective attention. *Visual Cognition*, 2:235–267.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dinh Dien, Kiem Hoang, and Nguyen van Toan. 2001. Vietnamese word segmentation. In *Natural Language Processing Pacific Rim Symposium*, pages 749–756.

Quang Thang Dinh, Hong Phuong Le, Thi Minh Huyen Nguyen, Cam Tu Nguyen, Mathias Rossignol, and Xuan Luong Vu. 2008. Word segmentation of Vietnamese texts: A comparison of approaches. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Tuyen Thi-Thanh Do. 2013. Building a vietnamese lexicon ontology for syntactic parsing and document

annotation. In *Proceedings of International Conference on Information Integration and Web-Based Applications & Services*, IIWAS '13, page 619–623, New York, NY, USA. Association for Computing Machinery.

Danh Thanh Do-Hunville and Huy Linh Dao. 2018. Transcategoriality and isolating languages: The case of Vietnamese. *Cognitive Linguistic Studies*, 5(1):8–38.

Mary S. Erbaugh. 2022. The acquisition of mandarin. *The Crosslinguistic Study of Language Acquisition*.

Nguyen Thien Giap. 2008. *Từ vựng học tiếng Việt (Vietnamese Lexicology)*. Vietnam Education Publishing House, Hanoi, Vietnam.

Nguyen Thien Giap. 2011. *Vấn đề "từ" trong tiếng Việt (The issue of "word" in Vietnamese*. Vietnam Education Publishing House, Hanoi, Vietnam.

David Gil. 2006. *Early Human Language Was Isolating-Monocategorial-Associational*, pages 91–98. World Scientific.

David Gil. 2008. How complex are isolating languages? In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity: Typology, Contact, Change*, volume 94 of *Studies in Language Companion Series*, pages 109–131. John Benjamins.

Cao Xuan Hao. 1998. *Tiếng Việt: Mấy vấn đề ngữ âm – ngữ pháp – ngữ nghĩa (Vietnamese: Selected Issues in Phonetics, Grammar, and Semantics)*. Vietnam Education Publishing House, Hanoi, Vietnam.

Nguyen Van Hiep. 2020. Heteroglossia: Another sfg-based approach to treatment of word order as a means for expressing modality in vietnamese. *VNU Journal of Foreign Studies*.

Nghia Hieu Nguyen, Dat Tien Nguyen, and Ngan Luu-Thuy Nguyen. 2025. Vietnamese words are not constructed from syllables: Rethinking the role of word segmentation in natural language processing for vietnamese texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):24069–24077.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Ho Le. 2003. *Vấn đề cấu tạo từ trong tiếng Việt hiện đại (Word Formation in Modern Vietnamese)*. Social Sciences Publishing House, Ha Noi, Vietnam.

Khai Le-Duc, David Thulke, Hung-Phong Tran, Long Vo-Dang, Khai-Nguyen Nguyen, Truong-Son Hy, and Ralf Schlüter. 2025. Medical spoken named entity recognition. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 724–783, Albuquerque, New Mexico. Association for Computational Linguistics.

Jo-Wang Lin. 2006. Time in a language without tense: The case of chinese. *J. Semant.*, 23:1–53.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yongxian Luo. 2020. Morphology in kra-dai languages. *Oxford Research Encyclopedia of Linguistics*.

Anh-Duong Nguyen, Kiem-Hieu Nguyen, and Van-Vi Ngo. 2019. Neural sequence labeling for vietnamese pos tagging and ner. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–5.

Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Le-Minh Nguyen, and Quang-Thuy Ha. 2006. Vietnamese word segmentation with CRFs and SVMs: An investigation. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 215–222, Huazhong Normal University, Wuhan, China. Tsinghua University Press.

Dat Quoc Nguyen, Dai Quoc Nguyen, Son Bao Pham, Phuong-Thai Nguyen, and Minh Le Nguyen. 2014. From treebank conversion to automatic dependency parsing for vietnamese. In *Natural Language Processing and Information Systems*, pages 196–207, Cham. Springer International Publishing.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Hien D. Nguyen, Tai Huynh, Suong N. Hoang, Vuong T. Pham, and Ivan Zelinka. 2020. Language-oriented sentiment analysis based on the grammar structure and improved self-attention network. In *Proceedings of the 15th International Conference on Evaluation of Novel Approaches to Software Engineering - ENASE*, pages 339–346. INSTICC, SciTePress.

Kiet Van Nguyen, Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018a. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 572–583, Cham. Springer International Publishing.

Phuong-Thai Nguyen, Xuan-Luong Vu, Thi-Minh-Huyen Nguyen, Van-Hiep Nguyen, and Hong-Phuong Le. 2009. Building a large syntactically-annotated corpus of Vietnamese. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 182–185, Suntec, Singapore. Association for Computational Linguistics.

Tuan-Phong Nguyen and Anh-Cuong Le. 2016. A hybrid approach to vietnamese word segmentation. *2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, pages 114–119.

Van-Nhat Nguyen, Nguyen Ha Thanh, Dinh-Hieu Vo, and Le-Minh Nguyen. 2018b. Relation extraction in vietnamese text via piecewise convolution neural network with word-level attention. *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 99–103.

Dmitry Nikolaev and Sebastian Padó. 2023. The universe of utterances according to BERT. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 99–105, Nancy, France. Association for Computational Linguistics.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does BERT rediscover a classical NLP pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Trang Phan and Eric T. Lander. 2015. Vietnamese and the np/dp parameter. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 60:391 – 415.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Virach Sornlertlamvanich and Wantenee Pantachat. 1993. Information-based language analysis for thai. *Asean Journal on Science and Technology for Development*, 10:181–196.

Maosong Sun, Jingyang Li, Zhipeng Guo, Yu Zhao, Yabin Zheng, Xiance Si, and Zhiyuan Liu. 2016. Thuctc: An efficient chinese text classifier. Natural Language Processing Lab, Tsinghua University.

Minna Suni. 2007. Awareness of second language inflectional morphology: A case study on finnish as a second language. *Acta Linguistica Hungarica*, 54:217–235.

Nguyen Kim Than. 1997. *Nghiên cứu về ngữ pháp tiếng Việt (Studies on Vietnamese Grammar)*. Vietnam Education Publishing House, Hanoi, Vietnam.

Nguyen Luong Tran, Duong Le, and Dat Quoc Nguyen. 2022. Bartpho: Pre-trained sequence-to-sequence models for vietnamese.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 named entity recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2146–2153, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Duy Vu-Tran, Phu-Thinh Pham, Duc Do, An-Vinh Luong, and Dien Dinh. 2022. Integrating label attention into CRF-based Vietnamese constituency parser. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 1–9, Manila, Philippines. Association for Computational Linguistics.

Dan Xiong, Qin Lu, Fengju Lo, Dingxu Shi, and Tinshing Chiu. 2015. Morpheme inversion in disyllabic compounds—cases in chinese diachronic corpora.

Xiang Xu, Wanxiang Shi, Yuanbin Luo, Zhi Gong, and Yuantao Chen. 2025. Ulner: A spoken-style chinese named entity recognition dataset for urban life domain. In *Proceedings of the 2024 International Symposium on AI and Cybersecurity*, ISAICS '24, page 45–49, New York, NY, USA. Association for Computing Machinery.

Liulin Zhang. 2023. Has chinese always been an analytic language? effects of writing on language evolution. *Language and Semiotic Studies*, 9:576 – 597.

Jiaomei Zhou and Zhiying Liu. 2022. Semantic similarity of inverse morpheme words based on word embedding. *Int. J. Asian Lang. Process.*, 31:2150006:1–2150006:13.

A Additional Analysis of Structural Bias Hyperparameters

The Structural Attention Matrix encodes inductive biases via two scalar parameters: α for intra-compound links and β for inter-compound links, applied to layers $1-2$ of the encoder. To systematically understand how these parameters affect the model, we perform a grid search over $\alpha \in [0, 5]$ and $\beta \in [-5, 0]$ with step size 0.5, fixing $\gamma = \delta = 0$. For each (α, β), we compute the attention maps of the base model A_{orig} (without bias) and the biased model A_{bias} on 10,000 sentences from the Vietnamese Book Corpus and measure:

$$C = \cos(A_{\text{orig}}, A_{\text{bias}}) \quad (1)$$

$$\|\Delta\|_F = \|A_{\text{bias}} - A_{\text{orig}}\|_F \quad (2)$$

Figure 3 shows cosine similarity and Frobenius norm over this grid. Several trends emerge:

Stability of pretrained geometry. Cosine similarity is strikingly high across most of the grid: $C \geq 0.98$ except at extreme values ($\alpha \gtrsim 4.5$, $|\beta| \gtrsim 4.5$). This indicates that, for moderate bias magnitudes, the structural prior acts as a gentle perturbation rather than rewriting the attention patterns learned during pretraining. In other words, the model retains its original inductive biases and only slightly reweights connections inside versus across compounds.

Magnitude of perturbation. The Frobenius norm $\|\Delta\|_F$ provides a complementary view of the absolute change in attention mass. As $|\alpha|$ or $|\beta|$ increases, $\|\Delta\|_F$ grows monotonically from values near 0 to about 2.48. There is a clear transition band around $|\alpha|, |\beta| \approx 2.0$: for $|\alpha|, |\beta| \lesssim 2.0$, the perturbation remains modest ($\|\Delta\|_F \lesssim 0.6$); beyond this region, the bias begins to dominate the original distribution ($\|\Delta\|_F \gtrsim 1.0$), signalling a regime where the structural prior can substantially overwrite pretrained behaviour.

Chosen operating point. Our main-setting ($\alpha = 0.5, \beta = -0.3$) lies deep inside the low-perturbation regime, achieving $C \approx 0.9995$ and $\|\Delta\|_F \approx 0.1447$. This reflects a conservative design: the bias is strong enough to encode a clear preference for intra-compound attention and a mild penalty for cross-compound links, yet weak enough to function as a soft guidance signal. Empirically, this configuration achieves the best trade-off between accuracy and representation stability on Vietnamese syntactic tasks (Table 1).

Asymmetry between α and β. The heatmaps also reveal an asymmetry: increasing α (rewarding intra-compound edges) tends to be less disruptive than aggressively decreasing β (penalizing cross-compound edges). Intuitively, rewarding within-compound connections preserves most of the original attentional mass while biasing it toward linguistically plausible spans, whereas strongly suppressing cross-compound edges can force attention to ignore potentially useful long-distance dependencies. This suggests that future tuning should prioritize moderate positive α and more conservative negative β values.

Implications for transfer and deployment. The smooth landscape of both metrics implies that the model is robust to small variations in (α, β) within the low-perturbation region. Practitioners can therefore treat small-magnitude biases (e.g., $\alpha \in [0.25, 1.0]$, $\beta \in [-1.0, 0]$) as safe defaults and adjust them to trade off structural strength against faithfulness to the pretrained encoder. This is particularly useful when transferring the method to other isolating languages, where one might prefer to start with conservative settings and gradually increase the bias based on downstream validation performance.

KL Divergence Analysis. To complement cosine similarity and Frobenius norm with a more sensitive distributional measure, we compute the per-entry KL divergence between A_{orig} and A_{bias} on the same 10,000 Vietnamese Book Corpus sentences ($\alpha = 0.5$, $\beta = -0.3$, $\gamma = \delta = 0$, all heads in Layers 1–2). As shown in Table 4, mean KL divergence remains on the order of 10^{-3} across both layers, consistently with the near-unity cosine similarities and small Frobenius norms reported above, further confirming that the structural prior introduces only minimal probabilistic divergence from the pretrained attention distributions.

Table 4: Per-entry KL divergence between pre- and post-bias attention distributions on Layers 1–2 of HuTieuBERT (10,000 Vietnamese Book Corpus sentences; $\alpha = 0.5$, $\beta = -0.3$, $\gamma = \delta = 0$, all heads).

Metric	Layer 1	Layer 2
Mean	1.351×10^{-3}	1.379×10^{-3}
Std	1.190×10^{-3}	9.670×10^{-4}

Overall, the ablation confirms that there is a broad regime in which structural guidance mean-

ingfully shapes attention while preserving the core representational geometry, and that HuTieuBERT operates well inside this stable region. This conclusion is further corroborated by the KL divergence analysis, where mean divergence values on the order of 10^{-3} across Layers 1–2 indicate that the post-bias attention distributions remain probabilistically close to their pretrained counterparts, providing converging evidence across all three metrics that the structural prior acts as a soft and stable inductive bias rather than a disruptive reweighting of the learned attention geometry.

B Illustrative Case: Structural Attention Bias on HuTieuBERT and MAChineseBERT Layers 1–2 full heads

We conduct an illustrative case study on two example sentences: Vietnamese *Tôi học ở Thành phố Hồ Chí Minh* ("I study in Ho Chi Minh City") and Chinese 美国弗州5.9级地震 ("5.9 magnitude earthquake in Virginia, USA"). Attention heatmaps are visualized on HuTieuBERT and MAChineseBERT to examine differences between biased and unbiased attention mechanisms, revealing how structural bias affects morpheme-aware processing. Syllable are color-coded by morpheme position: B (green), E (blue), M (yellow), S (black).

Attention weights are averaged across all 12 heads in Layers 1–2 with bias parameters $\alpha = 0.5$, $\beta = -0.3$, $\gamma = \delta = 0.0$. Table 5 reports statistical metrics: correlation C, Frobenius norm $\|\Delta\|_F$, standard deviation (StdDiff), and maximum absolute difference (MaxAbsDiff).

Table 5: Statistical comparison for two example sentences (Figs. 4 and 5), comparing pre- and post-bias mean attention averaged over all heads in Layers 1–2.

HuTieuBERT				
Layers	C	$\|\Delta\|_F$	StdDiff	MaxAbsDiff
1	0.9998	0.0399	0.0039	0.0165
2	0.9993	0.0565	0.0056	0.0265
MAChineseBERT				
Layers	C	$\|\Delta\|_F$	StdDiff	MaxAbsDiff
1	0.9979	0.1221	0.0102	0.0457
2	0.9984	0.1096	0.0091	0.0446

Figures 4 and 5 visualize attention heatmaps comparing WITH-BIAS, NO-BIAS, and their DIFFERENCE. High correlation ($C > 0.997$) and small differences ($\|\Delta\|_F < 0.13$, StdDiff < 0.011) confirm that structural bias subtly modulates attention toward morphological boundaries while preserving overall patterns and model stability.

C Tokenizer with BMES Examples

Our tokenization approach for both Vietnamese (HuTieuBERT) and Chinese (MAChineseBERT) employs fine-grained tokenization with BMES tags to mark word boundaries. Unlike PhoBERT's underscore-concatenated tokens (e.g., *sinh_viên*), we treat each syllable separately in Vietnamese (*sinh* with B tag, *viên* with E tag) and each character in Chinese, while preserving word-level information through BMES labels (Begin, Middle, End, Single). This design enables construction of structural bias matrices that enhance both intra-word and inter-word attention patterns, providing richer linguistic structure compared to treating multi-syllable words as atomic tokens (Table 6).

We attribute this gap to a vocabulary mismatch: PhoBERT/HuTieuBERT's BPE vocabulary, built from Vietnamese-only text, leaves English and Latin-derived medical terms largely out-of-vocabulary, whereas XLM-R's multilingual pretraining provides direct biomedical terminology coverage. As shown in Table 7, PhoBERT/HuTieuBERT fragments *l-carnitine* into meaningless units while XLM-R preserves recognizable subwords, and similarly retains the meaningful prefix *beta-* in *betacarotene*. Such fragmentation directly weakens NER embeddings for entity spans across broken subwords—an inherent limitation of monolingual initialization rather than a fine-tuning artifact.

D Illustrative Case: Structural Attention Bias in Example Vietnamese and Chinese Sentences

Figures 6 and 7 visualize the structural attention bias patterns (with $\alpha = 0.5$, $\beta = -0.3$, $\gamma = 0$, $\delta = 0$) applied to example sentences in Vietnamese and Chinese, respectively, where red values indicate increased intra-phrase attention and dark blue values indicate decreased inter-phrase attention. The diagonal red blocks correspond to within-compound tokens receiving positive bias (α), while off-diagonal blue regions reflect the penalty (β) for cross-compound attention. This fixed prior is then added to attention scores before softmax.

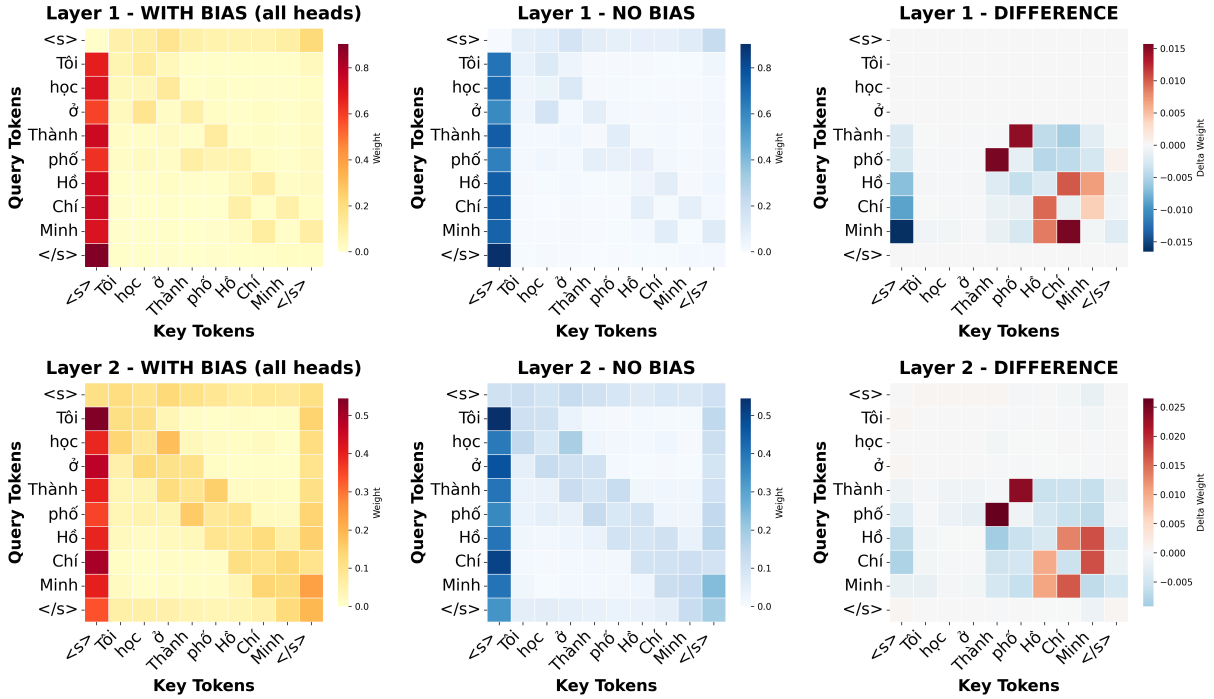

Figure 4: Attention weight heatmaps for HuTieuBERT processing Vietnamese sentence *Tôi học ở Thành phố Hồ Chí Minh* ("I study in Ho Chi Minh City"). across Layers 1–2 full heads. Each row compares: (a) WITH morpheme-aware BIAS, (b) NO BIAS baseline, (c) DIFFERENCE map. Red indicates increased attention, blue indicates decreased attention. The bias reshapes attention patterns to align with Vietnamese morphological boundaries.

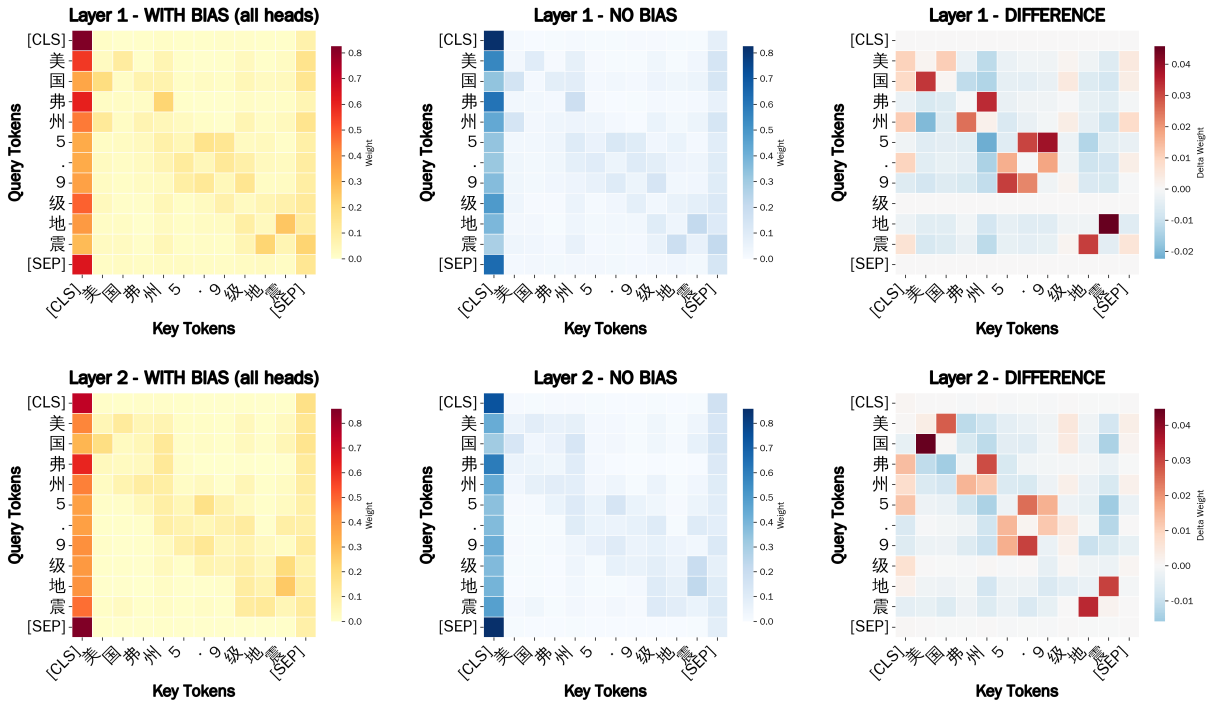

Figure 5: Attention weight heatmaps for MAChineseBERT processing Chinese sentence 美国弗州*5.9*级地震 ("5.9 magnitude earthquake in Virginia, USA") across Layers 1-2. Each row compares: (a) WITH morpheme-aware BIAS, (b) NO BIAS baseline, (c) DIFFERENCE map. Red indicates increased attention, blue indicates decreased attention. The bias modulates attention distribution around Chinese character-word boundaries.

Table 6: Example of Sentence and BMES Tokenization with Color-coded Tags

Sentence	Token with BMES tag
Tôi là sinh viên đại học Bách khoa	[“<s>”, “Tôi”, “là”, “sinh”, “viên”, “đại”, “học”, “Bách”, “khoa”, “<\s>”] [“S”, “S”, “S”, “B”, “E”, “B”, “E”, “B”, “E”, “S”]
Lập trình viên Python	[“<s>”, “lập”, “trì@@”, “nh@@” “viên”, “py@@”, “thon@@”, “<\s>”] [“S”, “B”, “M”, “M”, “E”, “S”, “S”, “S”]
Di tích này cần được khảo sát	[“<s>”, “di”, “tích”, “này”, “cần”, “được”, “khảo”, “sát”, “<\s>”] [“S”, “B”, “E”, “S”, “S”, “S”, “B”, “E”, “S”,]
甘肃临夏6.2级地震	[“<s>”, “甘”, “肃”, “临”, “夏”, “6”, “.”, “2”, “级”, “地”, “震”, “<\s>”] [“S”, “B”, “E”, “B”, “E”, “B”, “M”, “E”, “S”, “B”, “E”, “S”]
天安门国庆庆典	[“<s>”, “天”, “安”, “门”, “国”, “庆”, “庆”, “典”, “<\s>”] [“S”, “B”, “M”, “E”, “B”, “E”, “B”, “E”, “S”]
美国弗州5.9级地震	[“<s>”, “美”, “国”, “弗”, “州”, “5”, “.”, “9”, “级”, “地”, “震”, “<\s>”] [“S”, “B”, “E”, “B”, “E”, “B”, “M”, “E”, “S”, “B”, “E”, “S”]

Table 7: Subword tokenization of English medical terms under PhoBERT/HuTieuBERT versus XLM-R. Fewer tokens indicate more coherent segmentation.

Input term	HuTieuBERT/PhoBERT tokens	#	XLM-R tokens	#
l-carnitine fumarate	l-@@; car@@; nit@@; ine; f@@; um@@; ar@@; ate@@	8	_l; -; car; ni; tine; _fumar; ate	7
betacarotene	bet@@; ac@@; aro@@; ten@@; e@@	5	_beta; car; o; tene	4
phytoncide	phy@@; ton@@; c@@; ide	4	_; phy; ton; cide	4

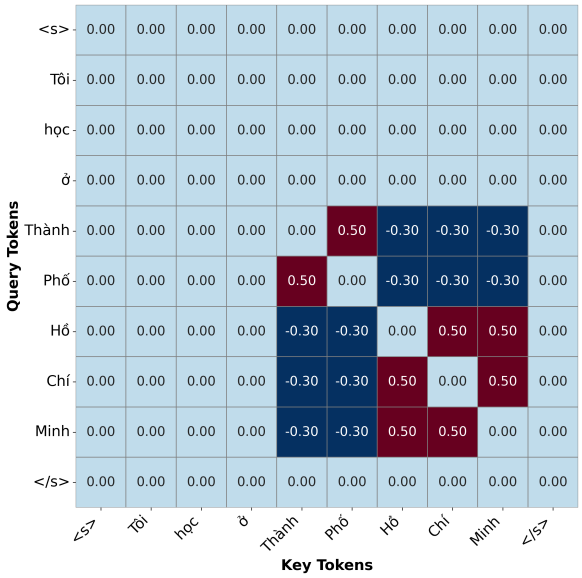

Figure 6: Structural Attention Bias of the sentence *Tôi học ở Thành Phố Hồ Chí Minh* (“I study in Ho Chi Minh City”) in Vietnamese.

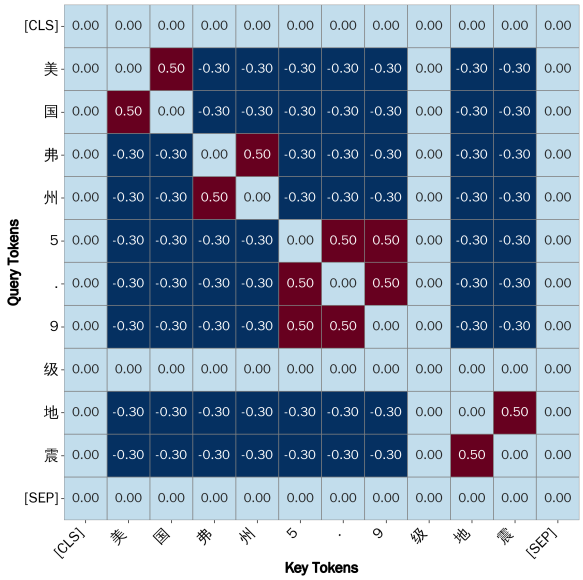

Figure 7: Structural Attention Bias of the sentence 美国弗州*5.9*级地震 (“5.9 magnitude earthquake in Virginia, USA”) in Chinese.