Sentence-Aware Bahnaric-Vietnamese Lexical Mapping with Contrastive Contextual Representations

Thi Ty Nguyen[1], **Phat T. Tran-Truong**[2], **Long S. T. Nguyen**[2], **Tan Sang Nguyen**[3], **Tho T. Quan**[2]
[1]RMIT University, Vietnam
[2]URA Research Group, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Vietnam
[3]National University of Singapore, Singapore
s4155541@student.rmit.edu.au, phatttt@hcmut.edu.vn, long.nguyencse2023@hcmut.edu.vn, e1583535@u.nus.edu, qttho@hcmut.edu.vn

Abstract

Underserved and extremely low-resource languages challenge current language technologies, especially when lexical borrowing and synonymy undermine exact-match assumptions. We study Bahnaric-Vietnamese lexical mapping as a step toward meaning-preserving sentence translation. Unlike prior work based on static embeddings and Mean Squared Error (MSE) alignment, we learn sentence-aware word representations with a small multilingual transformer pretrained on Vietnamese, adapt it with Low-rank adaptation (LoRA) for parameter efficiency, and align Bahnaric-Vietnamese pairs using a two-layer projection trained with InfoNCE contrastive loss. We exploit a new community-sourced lexicon of approximately 10,000 Bahnaric-Vietnamese pairs collected with local partners, capturing one-to-one, one-to-many, and many-to-one anchor relations as well as extensive lexical borrowing. Experiments evaluate retrieval-style alignment with Precision at K (P@K) and Mean Reciprocal Rank (MRR), as well as sentence translation using top-1 accuracy, Bilingual Evaluation Understudy (BLEU), Character n-gram F-score (ChrF), and embedding-based BERTScore. We also qualitatively analyze cases where n-gram metrics under-credit semantically adequate outputs in synonym-rich settings, and our ablation analysis shows that InfoNCE contrastive training dramatically outperforms MSE regression. On the $\sim$1k lexicon, our best model attains P@1 ≈ 0.53 and MRR ≈ 0.62, substantially improving over a static-embedding MSE baseline, while on the richer $\sim$10k community lexicon it reaches comparable sentence-level top-1 accuracy and BERTScore F1 despite slightly lower BLEU and chrF, highlighting both the benefits of the expanded resource and the remaining challenges of synonym-rich, low-frequency vocabulary.

1 Introduction

Language technologies continue to underserve communities whose languages lack large corpora, standardized orthographies, or widely available benchmarks. Bahnaric, which is an Austroasiatic language spoken by ethnic minority communities in Vietnam, is emblematic of this gap: available lexical resources are scarce, borrowing from Vietnamese is extensive, and many-to-one or one-to-many synonym relations are common. These properties break exact-match assumptions in conventional word mapping and sentence translation pipelines, and they complicate evaluation regimes that reward surface overlap over meaning preservation (Vo et al. 2024).

Prior work on Bahnaric-Vietnamese lexical mapping has primarily used static word embeddings aligned with regression losses (e.g., MSE) (La et al. 2023), which (i) ignore the role of sentential context in disambiguating polysemy and borrowings, and (ii) underperform when multiple target synonyms are equally acceptable. In contrast, our end goal is meaning-preserving *sentence* translation, where each token should be represented in its sentence context rather than as a single static vector.

This paper makes two practical advances toward that goal. First, on the modeling side, we learn *sentence-aware* word representations with a compact multilingual transformer pretrained on Vietnamese, and align Bahnaric-Vietnamese pairs via a two-layer projection trained with an InfoNCE contrastive objective. We adopt LoRA for parameter-efficient adaptation, enabling experimentation on modest computation while retaining competitive quality. Second, on the data side, we leverage a new community-sourced lexicon of 10,000 Bahnaric-Vietnamese pairs (collected with local partners), which captures one-to-one, one-to-many, and many-to-one anchor relations alongside systematic borrowing and close Subject-Verb-Object (SVO) syntactic parallels. This lexicon supports both supervised alignment and more realistic evaluation settings where multiple reference translations are valid.

We evaluate our approach through a three-stage pipeline: (i) sentence-level contrastive fine-tuning (reporting train and test loss across 10, 20, and 50 epochs and comparing InfoNCE against an MSE objective), (ii) retrieval-style bilingual word alignment (reporting P@1, P@5, and MRR on both the $\sim$1k and $\sim$10k lexicons), and (iii) sentence translation suitability via embedding-based retrieval (reporting top-1 accuracy, BLEU, chrF, and BERTScore F1 (Zhang et al. 2020), which provides a more meaning-aware assessment of adequacy). While human evaluation would be valuable for assessing adequacy in this synonym-rich, low-resource setting, we do not conduct a formal human study in this work due to the limited availability of native Bahnaric-Vietnamese bilingual speakers and resource constraints. We therefore rely on automatic metrics and qualitative analysis, and explicitly acknowledge the absence of human evaluation as a limitation. Our current analysis shows that

InfoNCE-based sentence representations consistently outperform static MSE-style regressors: MSE collapses to near-zero loss yet yields almost no usable retrieval, whereas InfoNCE improves monotonically with more epochs and supports robust word- and sentence-level alignment even on the noisier 10k lexicon. At the same time, n-gram metrics often under-credit valid synonym substitutions, orthographic variants, and paraphrases in this low-resource setting, whereas BERTScore more faithfully tracks semantic adequacy.

Contributions. This work makes three contributions toward meaning-preserving Bahnaric-Vietnamese translation for an underserved language.

- On the modeling side, we propose a sentence-aware alignment pipeline that combines a Vietnamese-pretrained multilingual encoder, LoRA-based contrastive fine-tuning with InfoNCE, and supervised Kabsch alignment over anchor lexicons, and we use it to perform cosine-based sentence retrieval.
- On the data and analysis side, we leverage and characterize a new 10k community-sourced Bahnaric-Vietnamese lexicon, showing how its one-to-many and many-to-one mappings, expressive vocabulary, and noisy orthography make alignment more realistic but also more challenging than the earlier 1k seed lexicon, while still supporting effective supervision when paired with contrastive learning.
- On the evaluation side, we establish a three-stage protocol: sentence-level contrastive learning (with an ablation over InfoNCE vs. MSE and over training epochs), bilingual word retrieval (P@1, P@5, MRR on both 1k and 10k lexicons), and sentence-level translation via retrieval (top-1 accuracy, BLEU, chrF, and BERTScore), supplemented with qualitative case studies and an ablation over anchor lexicon size and quality. We also detail our training hyperparameters to support reproducibility.

2 Related Works

Lexical Mapping with Static Embeddings. Early cross-lingual lexical mapping aligns monolingual word spaces learned with Skip-gram or CBOW (Mikolov et al. 2013) using linear maps and Procrustes-style objectives, sometimes with orthogonality constraints (Grossi, Lanzarotti, and Lin 2017; Joulin et al. 2018; Grave, Joulin, and Berthet 2019; Patra et al. 2019). This classical formulation underlies much of Bilingual Lexicon Induction (BLI), where a small seed dictionary is used to estimate a linear transformation that brings source and target embeddings into alignment (Artemova and Plank 2023). For low-resource languages, anchor-based training further improves robustness: instead of training monolingual embeddings from scratch, the target space is initialized with high-resource source vectors serving as anchors, yielding better bilingual consistency and stronger monolingual quality (Eder, Hangya, and Fraser 2021). In Bahnaric-Vietnamese, prior work follows this static paradigm, using MSE-based alignment (La et al. 2023), but such methods struggle with polysemy, borrowings, and synonym-rich mappings where multiple Vietnamese targets may be equally valid for a single Bahnaric form.

Contextual Representations for Bilingual Lexicon Induction. Multilingual Transformers generate context-sensitive token representations that better disambiguate senses and capture fine-grained semantic correspondences (Aldarmaki and Diab 2019). Massive multilingual models such as mBERT and XLM-R naturally induce partially aligned cross-lingual spaces through shared subword vocabularies and joint training on 100+ languages (Zhong et al. 2024), enabling BLI without explicit mapping functions and supporting unsupervised bitext mining for closely related dialects (Artemova and Plank 2023). Dictionary-driven approaches further exploit bilingual lexicons to iteratively transfer subword embeddings across languages, improving performance for typologically distant low resource languages such as Uyghur and Khmer (Sakajo et al. 2025). Building on this line of work, recent studies compare alignment and joint training strategies (Wang et al. 2020), while ProMap (El Mekki et al. 2023) leverages prompting for improved contextual mapping. Our work employs a compact Vietnamese-pretrained encoder to obtain sentence-aware representations tailored to Bahnaric-Vietnamese translation.

Contrastive Objectives & Regression Losses. Regression objectives (e.g., MSE) penalize absolute distance to a single target, which can be brittle when multiple targets are valid. In contrast, InfoNCE-style contrastive learning optimizes relative similarity: positives are pulled together while negatives are pushed apart (Oord, Li, and Vinyals 2018). This naturally accommodates one-to-many synonym sets by not forcing collapse onto a single point and by rewarding correct ranking among candidates (Liu et al. 2021). We therefore replace MSE with InfoNCE for alignment of contextual pairs and evaluate with retrieval metrics (P@K, MRR) that reflect ranked acceptability.

Parameter-Efficient Adaptation. Parameter-efficient fine-tuning techniques such as LoRA inject low-rank adapters into Transformer layers to adapt models with limited compute and without full weight updates (Hu et al. 2022). This is particularly pertinent for underserved languages where compute and data are scarce. We adopt LoRA to adapt a small multilingual encoder toward Bahnaric-Vietnamese alignment while keeping training efficient.

Low-Resource MT and Specialized LLMs. Broader work on low-resource natural language processing has explored complementary strategies for mitigating data scarcity. Machine translation (MT) is often used to synthesize monolingual or parallel data for underserved languages (Wei et al. 2022; Costa-Jussà et al. 2022; Artetxe et al. 2023; Nguyen et al. 2025), though such automatically generated text may fail to capture local linguistic nuances and can introduce hallucinations. A parallel trend develops specialized instruction-tuned LLMs for regional languages-such as SEA-LION for Southeast Asia (Ng et al. 2025), Atlas-Chat for Moroccan Arabic (Shang et al. 2025), and FarsInstruct

for Persian (Mokhtarabadi et al. 2025)-which improve downstream usability but still require substantial curated corpora and risk perpetuating biases from high-resource pretraining (Lai et al. 2023; Tao et al. 2024). Our work addresses a complementary dimension: fine-grained lexical and sentence-level alignment for Bahnaric-Vietnamese, where reliable community-sourced resources are particularly scarce.

Evaluation Under Synonymy and Borrowing. Standard exact-match or n-gram metrics (accuracy, BLEU, ChrF) can under-credit semantically adequate outputs when multiple synonyms or borrowed forms exist (Callison-Burch, Osborne, and Koehn 2006). Recent metrics that compare meaning in embedding space (e.g., BERTScore, COMET) offer better correlation with adequacy (Jauregi Unanue, Parnell, and Piccardi 2021), but require careful use under extreme low-resource regimes. Our study reports retrieval metrics (P@K, MRR) for alignment and standard translation metrics for comparability, and motivates synonym- and meaning-sensitive evaluation tailored to Bahnaric-Vietnamese where one-to-many and many-to-one anchors and systematic borrowing are common.

Regional and Community-Driven Resources. A core challenge for low-resource and Austroasiatic languages is the severe scarcity of digital resources (e.g., limited parallel corpora, sparse monolingual text, and the absence of comprehensive dictionaries), which constrains cross-lingual alignment and transfer learning efforts (Nguyen et al. 2025). Recent work on Southeast Asian languages also highlights issues of orthographic variation and community-sensitive data practices that affect benchmarking and fairness (Vo et al. 2024). The new 10,000-pair community-sourced Bahnaric-Vietnamese lexicon directly addresses these gaps by providing richer coverage and capturing one-to-one, one-to-many, and many-to-one anchors, including borrowed forms, enabling supervision and evaluation that more closely reflect real usage.

3 Dataset and Linguistic Characteristics

3.1 Lexicon Overview

Our study leverages two Bahnaric-Vietnamese lexicons:

- an earlier 1,000-pair seed lexicon used in prior work (La et al. 2023),
- a newly released 10,000-pair Bahnaric-Vietnamese lexicon, community-sourced from Bahnaric speakers in Vietnam's Central Highlands.

Each entry consists of a Bahnaric word and one or more Vietnamese equivalents, forming one-to-one, one-to-many, or many-to-one anchor mappings. This expanded lexicon enables both supervised alignment and more linguistically realistic evaluation, where multiple translations may be equally valid. Table 1 shows the comparison of the 1K and 10K Bahnaric-Vietnamese lexicons.

The 10K lexicon's larger size and finer-grained senses reduce out-of-vocabulary issues and support more precise lexical choices by adding ideophonic and descriptive items absent from the 1K lexicon. For instance, the 1K lexicon contains mostly basic, literal terms such as *'mi* "mưa" (rain). In contrast, the 10K lexicon includes far richer and more expressive forms, often ideophonic verbs that encode motion, texture, or sound with high granularity. Examples include:

- *'brẽk 'brẽk* - "tiếng cây sắp rẫy kêu răng rắc" (describing the *sharp cracking sound* of trees about to break);
- *'blit 'blat* - "ánh sáng lớn lóe lên" (referring to a *sudden, intense flash of light*);
- *tơ măn tơ 'dôh* - "bằng phẳng mênh mông" (depicting a *vast, perfectly flat open space*).

At the same time, the presence of long, paraphrastic Vietnamese glosses and highly expressive or affective Bahnaric forms introduces additional challenges. For example:

- *'buah* - "tiếc rẻ" (to feel deep regret);
- *'brơh* - "hãm hại" (to harm or sabotage someone);
- *hơ kar đum* - "da hồng hào" (describing a rosy skin tone).

reflect fine-grained semantic distinctions for which multiple Vietnamese interpretations may be equally valid. These nuances complicate token-level alignment and sentence-level lexical selection. Further difficulty arises from noisy orthography and compound Bahnaric entries (e.g., *'blơch 'blach*, *'bơ'blut*) paired with long descriptive Vietnamese glosses, which hinder subword segmentation and parameter sharing and can destabilize alignment and reduce the robustness of downstream MT models trained on this lexicon.

3.2 Linguistic Observations

Our new lexicon explicitly encodes synonym-rich correspondences: one Bahnaric word may map to several near-synonymous Vietnamese terms, as in *mi* ("rain") → "mưa," "mưa rào," or "mưa phùn," while multiple Bahnaric forms can also map to the same Vietnamese word due to dialectal or morphological variation. This many-to-one and one-to-many structure favors retrieval-based, ranking-aware evaluation (e.g., Precision@K, MRR) over strict exact-match metrics. In addition, Bahnaric shows high lexical borrowing from Vietnamese (e.g.,"Sơ Nông nghiệp păng Phat triên nông thôn" vs. "Sở Nông nghiệp và Phát triển nông thôn"), and the two languages share similar syntax (SVO order, adverb placement), as in "Adroi năm mẹ tơ tă" vs. "Trước khi đi, mẹ dặn:". These characteristics (synonymy, borrowing, and shared syntax-further) support the use of sentence-aware contextual encoders over static embeddings.

4 Methodology

We cast Bahnaric-Vietnamese mapping as supervised bilingual alignment rather than full sentence generation. A multilingual encoder with a two-layer projection head produces contextualized token representations, which are mean-pooled (optionally with IDF weighting) into sentence embeddings. On the Bahnaric side, these projected embeddings are mapped into the Vietnamese space using a rigid Kabsch transform (Glavaš and Vulić 2020) learned from anchor pairs (La et al. 2023). At evaluation time, we embed all Bahnaric and Vietnamese sentences in the test set and perform nearest-neighbor search in this shared space; for each Bahnaric sentence, we select the most semantically similar Vietnamese sentence using cosine similarity. Thus, our model

Table 1: Comparison of the 1K and 10K Bahnaric-Vietnamese lexicons.

Aspect	1K lexicon	10K lexicon
Size & coverage	Compact; core high-frequency nouns, verbs across many domains (nature, time, body, kinship, everyday actions).	Much larger; adds many rare, descriptive and expressive items.
Sense granularity	Mostly "dictionary-like" senses; one Bahnaric form → one Vietnamese word or short phrase.	Many fine-grained, nuanced senses and near-synonyms (e.g., different kinds of brightness, motion, texture).
Vietnamese (VN) translation (gloss) length & structure	VN side often single words or short noun phrases (e.g. *mưa*, *trăng tròn*, *tháng tư*).	VN side includes many longer phrase-like glosses and paraphrases (e.g. *ánh sáng lớn lóe lên*, *tiếng cây sắp rẫy kêu răng rắc*, *làm việc qua loa*).
Register, style & expressivity	More neutral, referential vocabulary (physical world, kinship, simple predicates).	Many expressive, colloquial and ideophonic forms (sound-symbolic verbs, manner adverbs, affective words).
Domain balance	Broad but relatively balanced semantic fields (environment, agriculture, body, daily life).	Skewed toward fine-grained descriptive and affective domains visible in the sample (motion, sound, texture, evaluation).
Orthography & consistency	More standardized VN forms; fewer highly idiosyncratic spellings or punctuation patterns.	More spelling variants, complex punctuation and multiword units (quotes, commas, internal spaces).

Table 2: Data sizes used in each experiment (numbers denote Bahnaric-Vietnamese pairs).

Task	Train	Test
Sentence-level alignment (fine-tuning)	51,930	–
1K lexicon alignment	832	214
10K lexicon alignment	8,285	2,073
Sentence-level retrieval evaluation	–	2,001

yields a *retrieved Vietnamese sentence* from a fixed index-not a generated translation-which makes the method suitable for tasks such as bilingual lexicon induction and semantic alignment, but not for open-ended machine translation in real-world settings.

4.1 Data and Task Setup

Table 2 summarizes all datasets used in our pipeline. We first fine-tune the multilingual encoder on 51,930 Bahnaric-Vietnamese sentence pairs to obtain sentence-aware representations. For supervised anchor-word alignment, we evaluate two lexicon settings:

- **1K lexicon:** 832 train pairs, 214 test pairs.
- **10K lexicon:** 8,285 train pairs, 2,073 test pairs.

For sentence-level retrieval (our translation prediction task), we evaluate on 2,001 held-out Bahnaric-Vietnamese sentence pairs using cosine nearest-neighbor search in the shared embedding space.

4.2 Training Setup and Hyperparameters

Unless otherwise noted, all experiments use the same encoder configuration and optimization setup summarized in Table 3. Both the Bahnaric (source) and Vietnamese (target) encoders are initialized from the same multilingual model and equipped with a two-layer projection head. We optimize a symmetric InfoNCE loss with in-batch negatives, train only the projection heads and LoRA adapters while freezing the remaining base encoder parameters, and keep the hyperparameters fixed across runs, varying only the number of training epochs (10, 20, 50) for our reported results.

4.3 Baseline

As a baseline, we reproduce a static-word setting similar to La et al. (2023). Bahnaric and Vietnamese Skip-gram embeddings are trained independently, then aligned using a three-layer feed-forward network optimized with MSE loss over paired word vectors. This approach (i) ignores sentence-level context and (ii) assumes a single target vector per Bahnaric word, making it brittle under synonymy, polysemy, and Vietnamese borrowings.

4.4 Proposed Pipeline

Our proposed pipeline consists of three components: (1) contextual fine-tuning to obtain sentence-aware representations, (2) anchor-word alignment via the Kabsch algorithm, and (3) cross-lingual sentence retrieval via cosine-based matching (Figure 1).

Contextual fine-tuning. For fine-tuning on Bahnaric-Vietnamese sentence data, our model replaces static word embeddings with a compact multilingual Transformer encoder pretrained on Vietnamese (not Bahnaric) (Reimers and Gurevych 2019). Bahnaric and Vietnamese sentences are tokenized using a shared subword vocabulary and encoded with the same encoder. A two-layer projection network then maps the hidden states from both languages into a shared embedding space. We optimize the model using the InfoNCE contrastive loss (Oord, Li, and Vinyals 2018; Liu et al. 2021), which prevents different sentences from collapsing into identical embeddings and encourages aligned sentence pairs to be close in the shared space. Let f_i denote the Bahnaric sentence embedding and g_i the corresponding Vietnamese embedding. The InfoNCE objective is

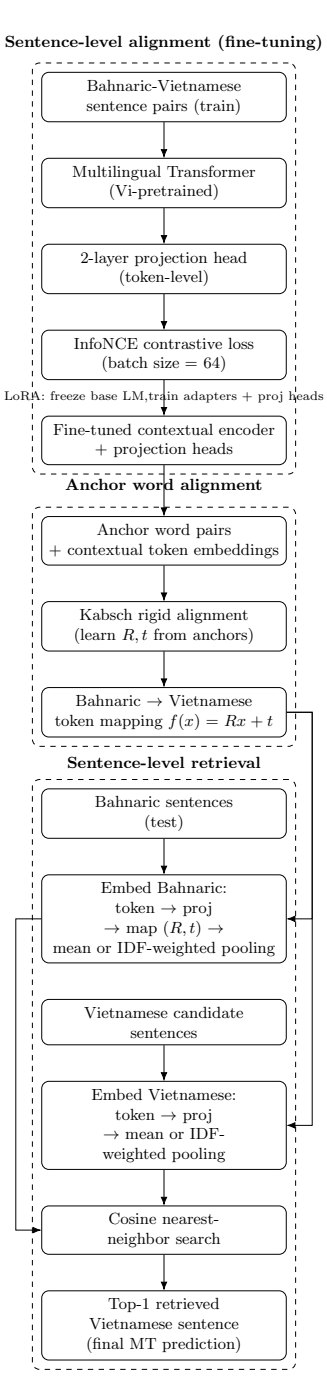

Figure 1: Overview of our methodology: (1) contextual fine-tuning for sentence-aware representations, (2) anchor word alignment via Kabsch, and (3) sentence translation via cosine retrieval.

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(f_i, g_i))}{\sum_j \exp(\text{sim}(f_i, g_j))}, \tag{1}$$

which forces the global structure of the space by ensuring that each f_i is closer to its true counterpart g_i than to all other negatives in the batch. To make the contrastive objective effective, we use a batch size of 64 to increase the number of negative samples. During fine-tuning, we apply LoRA, freezing all base model parameters and updating only the LoRA adapter weights and projection heads. This reduces training time while maintaining competitive performance.

Table 3: Key hyperparameters for sentence-level contrastive fine-tuning.

Hyperparameter	**Value**
Source & Target encoders	paraphrase-multilingual-MiniLM-L12-v2
Projection head	2-layer MLP (GELU, dropout 0.1)
Projection dimension	256
Batch size	64
Max sequence length (src & tgt)	256
Loss type	InfoNCE (symmetric, in-batch negatives)
Temperature τ	0.07
Optimizer	AdamW
Learning rate	2×10^{-4}
Epochs (reported runs)	10, 20, 50
Gradient clipping	max grad norm = 1.0
Random seed	42
Base encoders frozen	Yes (LoRA + projection only)
LoRA rank r	8
LoRA α	16
LoRA dropout	0.05
LoRA target modules	query, key, value, dense

Table 4: Evaluation metrics used across training stages.

Stage	**Metrics**
Sentence fine-tuning	Train loss, Test loss
Word alignment	P@1, P@5, MRR
Sentence translation	Top-1 accuracy, BLEU, chrF, BERTScore

Anchor Word Alignment. For anchor word alignment, we extract contextualized embeddings for each word pair using the fine-tuned model, then apply the unsupervised Kabsch algorithm to estimate a rigid transformation (R, t) that maps Bahnaric embeddings to Vietnamese. This follows the same alignment strategy as La et al. (2023).

Inference & Evaluation. For inference and evaluation, each test sentence is encoded and pooled into a sentence embedding. Translation is then performed via cosine-similarity retrieval over all candidate Vietnamese sentences, allowing us to evaluate the entire pipeline end-to-end. Table 4 summarizes evaluation metrics across the three stages.

Table 5: Training and test losses for different fine-tuning objectives and epoch settings.

Setting	Train Loss (avg)	Test Loss
10 epochs, MSE	0.000000	0.000000
10 epochs, InfoNCE	0.416097	1.157470
20 epochs, InfoNCE	0.203574	1.198847
50 epochs, InfoNCE	0.074888	1.055076

Table 6: Lexicon retrieval performance on the 1K Bahnaric-Vietnamese lexicon across epochs.

Epochs	P@1	P@5	MRR
10 epochs	0.092593	0.282407	0.166049
20 epochs	0.273148	0.546296	0.369676
50 epochs	0.527778	0.763889	0.624383

5 Experiments and Results

In this section, we evaluate our approach through three groups of experiments:

- sentence-level contrastive fine-tuning,
- bilingual word alignment using Kabsch, and
- sentence-level translation via embedding retrieval.

These correspond to the three stages of our pipeline and allow us to measure how improvements at the sentence level propagate to lexical alignment and full-sentence translation quality.

5.1 Sentence-Level Fine-Tuning

Table 5 summarizes the optimization behaviour of our sentence-level contrastive fine-tuning. We report average training loss and test loss under multiple epoch settings (10, 20, and 50 epochs), comparing InfoNCE against an MSE objective used in prior work (La et al. 2023).

We observe that InfoNCE loss decreases steadily with more training (from 0.42 at 10 epochs to 0.07 at 50 epochs), indicating better separation of positive and negative pairs. Test loss also improves at 50 epochs (approximately 1.06), showing more stable generalization. In contrast, the MSE objective collapses to near-zero loss on both train and test sets, offering no discriminative structure for retrieval, highlighting the importance of InfoNCE for effective sentence-level alignment.

5.2 Word Alignment Performance

We evaluate the effect of fine-tuning epoch count on bilingual lexicon induction using Precision@1, Precision@5, and MRR. Table 6 reports results for the old $\sim$1K lexicon, while Table 7 presents results using the expanded 10K lexicon.

For 1K lexicon (Table 6), performance improves sharply with more training: P@1 increases from 0.09 to 0.53, P@5 from 0.28 to 0.76, and MRR from 0.17 to 0.62 when moving from 10 to 50 epochs, showing strong gains from contextual contrastive learning on the smaller, cleaner lexicon.

Table 7: Lexicon retrieval performance on the new 10K Bahnaric-Vietnamese lexicon across epochs.

Epochs	P@1	P@5	MRR
10 epochs	0.011100	0.042954	0.021678
20 epochs	0.027510	0.074807	0.044442
50 epochs	0.083012	0.152510	0.108631

Table 8: Sentence-level translation performance using the 1K lexicon.

Epochs	Top-1 Acc	BLEU	chrF	BERTScore
10 epochs	0.2049	17.88	31.70	0.7385
20 epochs	0.2824	23.07	38.20	0.7666
50 epochs	0.3783	34.69	46.28	**0.7995**

For 10K lexicon (Table 7), improvements are smaller but consistent: P@1 rises from 0.01 to 0.08, P@5 from 0.04 to 0.15, and MRR from 0.02 to 0.11 across 10 to 50 epochs. The more modest gains reflect the 10K lexicon's higher lexical variability, noisy orthography, and many fine-grained or polysemous entries.

5.3 Sentence Translation Evaluation

We evaluate end-to-end sentence-level translation using cosine-similarity retrieval over the Vietnamese sentence index. Performance is measured with Top-1 accuracy, BLEU, chrF, and the recently added BERTScore F1, which provides a meaning-aware evaluation signal better suited for synonym-rich, low-resource settings. All experiments use identical model configurations unless noted: a MiniLM encoder with LoRA adapters, projection dimension 256, InfoNCE loss, temperature 0.07, batch size 64, and maximum sequence length 256 tokens. Table 8 reports results using the $\sim$1K lexicon for anchor alignment, while Table 9 uses the $\sim$10K community lexicon.

Across both lexicons, all metrics improve steadily with additional contrastive training. BERTScore mirrors this trend, rising from 0.74 to 0.80, indicating stronger semantic alignment even when n-gram overlap remains low. The $\sim$10K lexicon provides slightly better early-epoch performance, likely due to richer vocabulary coverage, but both settings converge to comparable results at 50 epochs.

5.4 Ablation Studies

We conduct ablations to isolate the contributions of (i) the contrastive objective, (ii) the amount of sentence-level fine-tuning, and (iii) the size and quality of the anchor lexicon. Unless otherwise noted, all experiments use the configuration in Section 4.2: MiniLM encoder, LoRA adapters, 256-d projection head, symmetric InfoNCE loss, batch size 64, and maximum sequence length 256.

InfoNCE vs. MSE. Prior Bahnaric-Vietnamese alignment relied on regression-style MSE losses (La et al. 2023). To assess the necessity of a contrastive objective, we directly

Table 9: Sentence-level translation performance using the 10K lexicon.

Epochs	Top-1 Acc	BLEU	chrF	BERTScore
10 epochs	0.2129	21.14	33.68	0.7450
20 epochs	0.2919	27.87	39.19	0.7729
50 epochs	0.3623	34.56	45.56	**0.7962**

Table 10: Ablation on the training objective (10 epochs). InfoNCE produces dramatically stronger retrieval than MSE under identical settings.

Metric	MSE 1K	InfoNCE 1K	MSE 10K	InfoNCE 10K
P@1	0.0014	0.0926	0.0010	0.0111
P@5	0.0019	0.2824	0.0034	0.0430
MRR	0.0016	0.1660	0.0020	0.0217
Top-1 Acc	0.0010	0.2049	0.0005	0.2129
BLEU	0.00	17.88	0.39	21.14
chrF	2.58	31.70	11.04	33.68
BERTScore	0.6085	0.7385	0.6053	0.7450

replace InfoNCE with MSE while keeping all other components identical. Table 5 shows that MSE quickly converges to near-zero loss but produces *severely degraded retrieval*. On the 1K lexicon, MSE yields P@1 $\approx$ 0.0014 and Top-1 accuracy $\approx$ 0.001, compared to 0.0926 and 0.2049 under InfoNCE. On the 10K lexicon, the gap is similar: MSE achieves negligible retrieval (P@1 $\approx$ 0.0010; Top-1 accuracy $\approx$ 0.0005), while InfoNCE attains P@1 = 0.0111 and Top-1 accuracy = 0.2129. BLEU and chrF under MSE collapse close to zero.

In contrast, InfoNCE maintains separation between positives and negatives, yielding substantial gains in P@K, MRR, Top-1 accuracy, BLEU, chrF, and BERTScore. These trends confirm that the contrastive objective (not the encoder alone) is essential for effective cross-lingual alignment.

Effect of Fine-Tuning Epochs. Tables 6-9 show that increasing the number of contrastive training epochs consistently improves both word- and sentence-level retrieval. On the 1K lexicon, P@1 rises from 0.09 (10 epochs) to 0.53 (50 epochs), and MRR from 0.17 to 0.62. Sentence-level Top-1 accuracy similarly increases from 0.20 to 0.38, with BLEU improving from 17.9 to 34.7 and BERTScore F1 from 0.74 to 0.80.

Although the 10K lexicon is noisier, the same monotonic pattern appears: P@1 increases from 0.011 to 0.083, Top-1 accuracy from 0.213 to 0.362, and BERTScore F1 from 0.745 to 0.796. No overfitting is observed within 50 epochs, suggesting that longer contrastive training robustly sharpens cross-lingual semantic structure.

Anchor Lexicon Size: 1K vs. 10K. We next compare alignment using the earlier 1K lexicon against the new 10K community lexicon. At the word level, the 1K lexicon yields substantially higher P@K and MRR because it is cleaner and dominated by one-to-one dictionary-style mappings. The 10K lexicon includes many expressive forms, paraphrastic glosses, and one-to-many or many-to-one relations, which reduce exact-match retrieval scores.

However, at the sentence level, the 10K lexicon provides clear early-epoch benefits. At 10 epochs, models aligned using the 10K lexicon achieve higher Top-1 accuracy (0.2129 vs. 0.2049), BLEU (21.14 vs. 17.88), and chrF (33.68 vs. 31.70). This indicates that broader anchor coverage yields a better global mapping for sentence representations. By 50 epochs, both lexicon settings converge to similar translation performance (Top-1 accuracy $\approx$ 0.36-0.38, BLEU $\approx$ 34-35, chrF $\approx$ 45-46, BERTScore F1 $\approx$ 0.80). These results support our claim that the expanded 10K lexicon offers a more realistic and challenging alignment scenario while still enabling strong sentence-level retrieval when combined with contrastive learning.

5.5 When n-gram Metrics Under-Credit Semantically Adequate Outputs

BLEU and chrF depend on exact n-gram matches, so they often penalize correct predictions under synonymy, orthographic variation, or paraphrasing, which are patterns common in Bahnaric-Vietnamese. BERTScore, which measures contextual semantic similarity (Zhang et al. 2020), is far more robust to such surface differences, though it can still vary with reference choice and domain (Hanna and Bojar 2021).

A clear example is synonym substitution: retrieving "mưa rào" when the reference is "mưa." BLEU drops due to token mismatch, while BERTScore assigns high similarity, producing the smoother improvements seen in Tables 8-9.

BLEU is also sensitive when the model predicts a standardized Vietnamese form but the reference uses nonstandard spelling, reordered constituents, or morphological variants (e.g., "điều chỉnh" vs. "chỉnh sửa"). BERTScore remains stable because it evaluates meaning, not surface form.

Paraphrastic glosses further highlight BLEU's limitations. Descriptive entries like

- *'brẽk 'brẽk* - "tiếng cây sắp rẫy kêu răng rắc" or
- *'blit 'blat* - "ánh sáng lớn lóe lên"

share little lexical overlap with retrieved sentences even when the meaning aligns. BLEU/chrF therefore drop, while BERTScore stays high (e.g., BERTScore F1 $\approx$ 0.80 at 50 epochs).

Finally, because our system performs sentence retrieval, not generation, it often returns valid paraphrases or more formal Vietnamese sentences. BLEU penalizes these forms, whereas BERTScore more accurately reflects their semantic adequacy.

5.6 Qualitative Sentence Translation Examples

Table 11 presents representative predictions after 50 epochs of contrastive training (InfoNCE, LoRA, MiniLM encoder, 10K lexicon). These examples illustrate where the retrieval-based system succeeds (ID 2, 4) and where it fails (ID 1, 3, 5).

Table 11: Example Bahnaric sentences and predicted Vietnamese translations from the proposed model after 50 training epochs.

ID	Bahnaric input	Predicted Vietnamese
1	Sở nông nghiệp păng tơ iung pơ lei đei xơ nong tổng hợp, pơ tâng uỷ ban nhân dân tỉnh, Bộ nông nghiệp păng tơ iung pơ lei adroi năr 25 rim khei (phara khei 12 pơ truh pơ tâng adroi năr 10) Păng pơ tâng đột xuất ah đei hơ 'nha.	Xét tặng danh hiệu "Nghệ nhân nhân dân", "Nghệ nhân ưu tú" trong lĩnh vực di sản văn hóa phi vật thể lần thứ Ba - Năm 2021 tỉnh Bình Định (sau đây gọi tắt là Hội đồng cấp tỉnh).
2	Bok nguyễn hoàng sâm - chủ tịch khul doanh nhân hơ muh tỉnh .	Ông Nguyễn Hoàng Sâm - Chủ tịch Hội Doanh nhân trẻ tỉnh.
3	wei năng rim năr; rim năr a thai ling lang năng xu tơ drong hoat động dih ka vă đei tơ drong xư lý ko tă ah tình huống krup oei lech. Năng xu mă tơ pă dơng tơ drong xa truh truh tơ drong hơ blih môi trường(mă loi, PH hơ duh mê).	Đồng thời, huyện cũng mở các lớp dạy nghề ngắn hạn, tổ chức tập huấn nâng cao kiến thức trồng trọt, chăn nuôi, hướng dẫn người dân áp dụng các tiến bộ khoa học kỹ thuật vào sản xuất, từng bước xóa đói, thoát nghèo hiệu quả.
4	Hăp tok ko jung truh dôm linh tơ plenh, ăm ko minh pah linh noh păng minh pah dôm xơ nglong a thai văr truh teh, noh hăp joa jâng tok.	Nó vươn cao tới tận các đạo binh trên trời, làm cho một phần đạo binh đó và một phần các tinh tú phải nhào xuống đất, rồi nó đạp chân lên.
5	Hăp hơ xoang khil pôm minh rơ wưh đek.	Đi bên chàng có khó khăn mấy em sẽ quyết vượt qua, sợ gì chứ.

These examples highlight three characteristics of our system: (i) sentence-aware representations enable recovery of fluent Vietnamese despite noisy Bahnaric orthography, (ii) retrieval naturally handles borrowed and near-synonymous Vietnamese forms without requiring generation, and (iii) the model remains robust across domains with long, complex sentences typical of real Bahnaric usage.

Evaluation limitations. Our evaluation relies on automatic metrics (P@K, MRR, BLEU, chrF, and BERTScore) and qualitative analysis. We do not conduct a formal human evaluation of adequacy or fluency, which is an important limitation given the synonym-rich and orthographically variable nature of Bahnaric-Vietnamese. Automatic metrics may under- or over-estimate perceived translation quality, particularly for low-frequency or highly expressive lexical items. Addressing this limitation through controlled human evaluation with native speakers remains an important direction for future work.

6 Conclusion and Future Work

We presented a sentence-aware Bahnaric-Vietnamese alignment pipeline combining a Vietnamese-pretrained multilingual encoder, InfoNCE contrastive learning, LoRA-based adaptation, and supervised Kabsch alignment. The model consistently outperforms a static-embedding MSE baseline: on the 1K lexicon it reaches P@1 ≈ 0.53, MRR ≈ 0.62, and up to 0.38 top-1 accuracy, 34.7 BLEU, 46.3 chrF, and 0.80 BERTScore F1. Despite greater orthographic noise, paraphrastic glosses, and one-to-many mappings, the 10K lexicon yields comparable sentence-level performance and stable gains with more training. Methodologically, contextual representations and contrastive training prove robust under synonymy and borrowing, while LoRA enables efficient adaptation with limited compute. Retrieval further accommodates multiple valid Vietnamese candidates without requiring a generative decoder. BERTScore complements n-gram metrics by better reflecting semantic adequacy, especially in cases where BLEU and chrF under-credit valid synonym or paraphrase matches.

Several challenges remain: the 10K lexicon introduces noise that depresses P@K and BLEU; fine-grained and low-frequency senses limit anchor reliability; and retrieval-based translation cannot produce unseen or compositional outputs, making performance dependent on the coverage and quality of the Vietnamese index. Future work includes: enhanced retrieval via paraphrase-augmented indexes or cross-encoder reranking; lexicon refinement, including normalization, sense disambiguation, and clustering; richer evaluation methodologies, including the design of human assessment protocols for adequacy and fluency with Bahnaric-Vietnamese speakers when such evaluation becomes feasible; progress toward generation through retrieval-augmented or lightweight generative models; and expanded community data across dialects, domains, and contextual uses.

Overall, sentence-aware contrastive alignment offers a practical step toward meaning-preserving Bahnaric-Vietnamese translation according to automatic evaluation metrics, while highlighting the linguistic complexity and resource constraints that motivate further modeling and community-driven data development.

Acknowledgments

We would like to thank Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, for supporting this study, and the URA Research Group at HCMUT for providing the 1K and 10K Bahnaric-Vietnamese lexicons and the approximately 52K parallel sentence pairs used in our experiments.

References

Aldarmaki, H.; and Diab, M. 2019. Context-Aware Cross-Lingual Mapping. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 3906–3911. Minneapolis, Minnesota: Association for Computational Linguistics.

Artemova, E.; and Plank, B. 2023. Low-resource Bilingual Dialect Lexicon Induction with Large Language Models. In Alumäe, T.; and Fishel, M., eds., *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 371–385. Tórshavn, Faroe Islands: University of Tartu Library.

Artetxe, M.; Goswami, V.; Bhosale, S.; Fan, A.; and Zettlemoyer, L. 2023. Revisiting Machine Translation for Cross-lingual Classification. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6489–6499. Singapore: Association for Computational Linguistics.

Callison-Burch, C.; Osborne, M.; and Koehn, P. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In McCarthy, D.; and Wintner, S., eds., *11th Conference of the European Chapter of the Association for Computational Linguistics*, 249–256. Trento, Italy: Association for Computational Linguistics.

Costa-Jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; et al. 2022. No language left behind: Scaling human-centered machine translation.

Eder, T.; Hangya, V.; and Fraser, A. 2021. Anchor-based bilingual word embeddings for low-resource languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 227–232.

El Mekki, A.; Abdul-Mageed, M.; Nagoudi, E. B.; Berrada, I.; and Khoumsi, A. 2023. ProMap: Effective bilingual lexicon induction via language model prompting. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 577–597.

Glavaš, G.; and Vulić, I. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 7548–7555.

Grave, E.; Joulin, A.; and Berthet, Q. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1880–1890. PMLR.

Grossi, G.; Lanzarotti, R.; and Lin, J. 2017. Orthogonal procrustes analysis for dictionary learning in sparse linear representation. *PloS one*, 12(1): e0169663.

Hanna, M.; and Bojar, O. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, 507–517.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jauregi Unanue, I.; Parnell, J.; and Piccardi, M. 2021. BERTTune: Fine-Tuning Neural Machine Translation with BERTScore. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 915–924. Online: Association for Computational Linguistics.

Joulin, A.; Bojanowski, P.; Mikolov, T.; Jégou, H.; and Grave, E. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2979–2984.

La, H. C.; Le, M. Q.; Tran, O. N.; Le, D. D.; Nguyen, D. Q.; Nguyen, S. T.; Tran, Q.; and Quan, T. T. 2023. Low-Rank Adaptation Approach for Vietnamese-Bahnaric Lexical Mapping from Non-Parallel Corpora. *VNUHCM Journal of Engineering and Technology*, 6(SI8): 19–32.

Lai, V. D.; Ngo, N.; Veyseh, A. P. B.; Mẫn, H.; Dernoncourt, F.; Bui, T.; and Nguyen, T. H. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the association for computational linguistics: EMNLP 2023*, 13171–13189.

Liu, C.; Fu, Y.; Xu, C.; Yang, S.; Li, J.; Wang, C.; and Zhang, L. 2021. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8635–8643.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mokhtarabadi, H.; Zamani, Z.; Maazallahi, A.; and Manshaei, M. H. 2025. Empowering Persian LLMs for Instruction Following: A Novel Dataset and Training Approach. In Hettiarachchi, H.; Ranasinghe, T.; Rayson, P.; Mitkov, R.; Gaber, M.; Premasiri, D.; Tan, F. A.; and Uyangodage, L., eds., *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, 31–67. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Ng, R.; Nguyen, T. N.; Huang, Y.; Tai, N. C.; Leong, W. Y.; Leong, W. Q.; Yong, X.; Ngui, J. G.; Susanto, Y.; Cheng, N.; et al. 2025. Sea-lion: Southeast asian languages in one network.

Nguyen, L.; Le, T.; Nguyen, H.; Vo, Q.; Nguyen, P.; and Quan, T. 2025. Serving the Underserved: Leveraging BART-Bahnar Language Model for Bahnaric-Vietnamese Translation. In Truong, S.; Putri, R. A.; Nguyen, D.; Wang, A.; Ho, D.; Oh, A.; and Koyejo, S., eds., *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, 32–41. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-242-8.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding.

Patra, B.; Moniz, J. R. A.; Garg, S.; Gormley, M. R.; and Neubig, G. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Sakajo, H.; Ide, Y.; Vasselli, J.; Sakai, Y.; Tian, Y.; Kamigaito, H.; and Watanabe, T. 2025. Dictionaries to the Rescue: Cross-Lingual Vocabulary Transfer for Low-Resource Languages Using Bilingual Dictionaries. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 25963–25976. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Shang, G.; Abdine, H.; Khoubrane, Y.; Mohamed, A.; Abbahaddou, Y.; Ennadir, S.; Momayiz, I.; Ren, X.; Moulines, E.; Nakov, P.; Vazirgiannis, M.; and Xing, E. 2025. Atlas-Chat: Adapting Large Language Models for Low-Resource Moroccan Arabic Dialect. In Hettiarachchi, H.; Ranasinghe, T.; Rayson, P.; Mitkov, R.; Gaber, M.; Premasiri, D.; Tan, F. A.; and Uyangodage, L., eds., *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, 9–30. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Tao, Y.; Viberg, O.; Baker, R. S.; and Kizilcec, R. F. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9): 346.

Vo, H. N. K.; Le, D. D.; Phan, T. M. D.; Nguyen, T. S.; Pham, Q. N.; Tran, N. O.; Nguyen, Q. D.; Vo, T. M. H.; and Quan, T. 2024. Revitalizing bahnaric language through neural machine translation: challenges, strategies, and promising outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23360–23368.

Wang, Z.; Xie, J.; Xu, R.; Yang, Y.; Neubig, G.; and Carbonell, J. G. 2020. Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Zhong, T.; Yang, Z.; Liu, Z.; Zhang, R.; Liu, Y.; Sun, H.; Pan, Y.; Li, Y.; Zhou, Y.; Jiang, H.; et al. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research.