MekongPhon: A Large-Scale Parallel IPA Corpus for Lao and Khmer

Ammon Shurtz, Christian Richardson, Stephen D. Richardson
Brigham Young University
Provo, UT
{acshurtz, richachr, srichardson}@byu.edu

Abstract

High-quality International Phonetic Alphabet (IPA) transcriptions are a foundational resource for speech and language technologies, yet existing tools for many low-resource languages remain limited in accuracy and scope. In this work, we present MekongPhon, a large-scale, high-quality parallel IPA corpus for Lao and Khmer. The corpus contains 1.3 million Khmer and 367 thousand Lao orthographic–IPA pairs, meticulously aligned and verified. When used to train Transformer-based sequence-to-sequence models, MekongPhon enables exceptionally accurate IPA generation, achieving under 2% Character Error Rate (CER) on held-out test sets. We further introduce linguistically informed Lao and Khmer transliteration tools that offer high-speed IPA conversion, outperforming Epitran by 6-71 CER points despite trading some accuracy for efficiency. All data, code, and pretrained models are publicly released to support future research and development in low-resource language technologies.

Keywords: Corpus Creation, Less-Resourced/Endangered Languages, Phonetic Databases, Phonology, Speech Resource/Database, Tools, Systems, Applications

1. Introduction

High-quality phonetic data, such as transcriptions in the International Phonetic Alphabet (IPA), is a foundational resource for speech and language technologies. It is the bridge between written text and spoken sound, making it a critical component for both Text-to-Speech (TTS) synthesis and Automatic Speech Recognition (ASR) (Cheng et al., 2024; Mortensen et al., 2018). Furthermore, IPA is the standard for phonetic and phonological analysis, allowing linguists to accurately document and compare sounds across languages (International Phonetic Association, 2020). While robust Grapheme-to-Phoneme (G2P) tools exist for high-resource languages, low-resource languages like Lao and Khmer are underserved. Existing tools for these languages often suffer from high error rates, lacking the quality and coverage needed to build effective downstream applications (Lin et al., 2022; Soky et al., 2016).

The need for automating this transcription process is significant, as Khmer and Lao are spoken by millions of first-language speakers (approximately 18M and 3.8M, respectively).[1] A popular, broad-coverage tool for this Grapheme-to-Phoneme (G2P) task is Epitran (Mortensen et al., 2018), which supports over 150 languages, including both Lao and Khmer. Despite this broad support, its performance on these specific languages is inadequate for downstream tasks. On our gold-standard test set (detailed in Section 4.1), Epitran yields a Character Error Rate (CER) of 105.57 for Khmer and 33.48 for Lao. A CER exceeding 100, as seen with Khmer, indicates that the number of IPA character errors (insertions, deletions, and substitutions) is greater than the total number of characters in the reference, rendering it highly unreliable.

This poor performance highlights a critical resource gap: the lack of large-scale, high-quality phonetic training data for Lao and Khmer. To address this gap, we introduce MekongPhon, a new, large-scale parallel phonetic resource. Our method leverages existing human-annotated IPA from historical dictionaries to create a high-precision seed lexicon. We then use this lexicon, structured in a computationally efficient trie, to convert a large, general-domain corpus. By filtering to keep only sentences that were completely converted (i.e., contain no unknown orthographic characters), we produced a high-fidelity parallel corpus of 1.3M Khmer and 367K Lao sentences.

In this work, we demonstrate the value of this new resource by comparing both rule-based scripts and neural Transformer models for G2P. We show that these models, trained on MekongPhon, dramatically outperform the Epitran baseline on our held-out test set. We make three primary contributions:

1. Large-scale Lao and Khmer parallel IPA corpora
2. Our improved, high-performance rule-based scripts
3. High-accuracy trained neural models

[1] https://www.ethnologue.com

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1650–1658
11–16 May 2026.

To ensure reproducibility and encourage future research, all resources—including the IPA lexicons, 1.3M-sentence Khmer and 367K-sentence Lao parallel corpora, our rule-based scripts, and the pre-trained neural models—are made freely available at https://github.com/byu-matrix-lab/MekongPhon.

2. Related Work

2.1. Phonetic Resources for Lao and Khmer

Several publicly available resources provide phonetic inventories or lexicons for Lao and Khmer. These include broad-coverage databases like PHOIBLE, which collates phonological inventories from descriptive grammars (Moran and McCloy, 2019), and crowd-sourced lexicons such as Wiktionary. More recently, the WikiPron project (Lee et al., 2020) has provided scraped phonetic data from Wiktionary for over 165 languages, including both Lao and Khmer.

While these resources are valuable, crowd-sourced data can suffer from inconsistencies. To build a high-precision seed lexicon for our corpus, we instead base our work on two comprehensive, linguist-curated resources: the 1977 Cambodian-English dictionary (Headley, 1977) and the 1972 Lao-English dictionary (Kerr, 1972). The critical advantage of these sources is that they contain **human-annotated IPA transcriptions** for every entry, providing a consistent, high-fidelity foundation that is ideal for training a robust G2P model.

2.2. Grapheme-to-Phoneme (G2P) Baselines

The most prominent G2P baseline for these languages is Epitran (Mortensen et al., 2018). For both Khmer and Lao, Epitran employs its "Simple Epitran" backend, which relies on a static map file to define direct mappings between orthographic strings and IPA strings.

However, we found these mappings to be insufficient for high-quality transliteration, especially for Khmer. A manual inspection of its map file (`epitran/data/map/khm-Khmr.csv`) reveals a critical omission: it includes mappings for the 33 consonants but **omits all vowel diacritics and dependent vowel symbols**. This flaw makes it fundamentally unable to transliterate the vast majority of Khmer words. The support for Lao is more complete, as its map file includes most letters and another file includes preprocessing rules, but it still suffers from coverage gaps and errors, as our evaluation in Section 4 demonstrates. This performance gap motivates our work in creating not only a large-scale corpus but also new, high-performance rule-based and neural G2P models.

3. MekongPhon Resource Creation

3.1. Seed Lexicon Curation

Our work is founded on a high-precision, **human-annotated** seed lexicon. We derived this lexicon from digitized dictionaries hosted by the SEALang Library[2], a database of South-East Asian language resources.

A critical preliminary step was the manual verification of transcription quality. We found that many resources on the SEALang site were automatically transliterated using a rudimentary rule-based system, which, for example, failed to disambiguate pronunciation based on a consonant's position (e.g., initial vs. final). To ensure quality, we exclusively selected resources with verified human-annotated IPA: the **1977 Cambodian-English dictionary** (Headley, 1977) for Khmer and the **1972 Lao-English dictionary** (Kerr, 1972) for Lao.

Data extraction presented a significant technical challenge. The SEALang Library project is unmaintained, and its web interface does not support large-scale queries. Furthermore, programmatically retrieving entries was hampered by incomplete metadata tags. To overcome these limitations, we developed a custom scraping solution. We first obtained comprehensive word lists from the library's frequency lists.[3] We then used these lists to systematically query the database API with the Python `requests` library, parsing the responses to extract entries *only* from our two selected high-quality dictionary editions.

This process yielded **8,026 Khmer word-IPA pairs** (from 17,920 words in the frequency list) and **6,039 Lao word-IPA pairs** (from 15,242 words). The Lao dictionary notably provides both tonal and non-tonal IPA representations for its entries, which we preserved. To create a complete lexicon for sentence-level G2P, we manually added mappings for all 10 orthographic numerals (0-9) and common punctuation marks. This final, curated lexicon formed the basis for our large-scale corpus expansion.

Seed Lexicon Qualitative Analysis We randomly sample 500 entries from the constructed seed lexicon and conduct a qualitative human evaluation. Each sampled entry is manually reviewed by two of the authors to verify that the orthographic form (Khmer or Lao) appropriately corresponds to

[2] http://sealang.net/library/ The SEALang homepage states that "all results are freely available for re-use under the Creative Commons open license."

[3] http://sealang.net/project/list/

Statistic	Khmer	Lao
Sentences	1,314,779	367,388
Unique Tokens	28,096	32,302
Avg. Tokens / Sent.	14.18	13.45
Avg. Length Sent. (Chars)	44.06	40.36

Table 1: MekongPhon corpus statistics for the Khmer and Lao parallel IPA data. The XLM-Roberta (Conneau et al., 2020) tokenizer was used as it supports both Lao and Khmer, which do not delimit words with spaces.

the IPA transcription provided by the source dictionary. Minor discrepancies in IPA representation are still classified as accurate, as some degree of phonetic variation is expected due to natural and dialectal variation in speech. Each entry is classified as either accurate or inaccurate.

We note that both authors are highly proficient second-language speakers rather than native speakers, and subtle phonetic judgments may therefore differ from native-speaker intuitions.

For Khmer, 4 of the 500 sampled entries were found to be inaccurate, resulting in 99.2% accuracy. Three of the four inaccurate entries involve a final consonant being transcribed with its inherent vowel retained, as though it were in onset position. For example, the word សាមញ្ញ is transcribed as /saamaɲɲeaʔ/ rather than /saamaɲ/, where the inherent vowel of the final ញ is incorrectly retained, treating it as an onset. The remaining error involved a missing a medial consonant.

For Lao, 14 of the 500 sampled entries were found to be inaccurate, resulting in 97.2% accuracy. The majority of errors (10 out of 14) involved incorrect vowel and/or final consonant sounds, such as ວັຍ being transcribed as /wiaːŋˈ/ rather than the correct vowel and final. Three further samples introduced a spurious syllable absent from the orthography; for instance, ຣັຊ was transcribed as /raːt\saˈdɔːn/, where the syllables saˈdɔːn have no orthographic basis. The remaining errors consisted of one entry transcribed as an entirely different word — ວັຕ rendered as /wiaːt\naːmˈ/ — and one case of an incorrect initial consonant.

3.2. Corpus Expansion and Filtering

To generate a large-scale parallel corpus, we first sourced monolingual data for both languages from the OPUS database (Tiedemann, 2012). We aggregated data from its largest available collections: NLLB (Schwenk et al., 2020; Fan et al., 2020), Paracrawl (Koehn, 2024), and OpenSubtitles (Lison and Tiedemann, 2016). This resulted in a source corpus of 8,962,874 unique Khmer sentences and 4,609,363 unique Lao sentences.

We then developed a high-speed, high-precision transliteration pipeline to convert this orthographic text. The seed lexicon (described in Section 3.1) was loaded into a trie (prefix tree), a data structure that enables highly efficient lookups. We applied a greedy, longest-match-first algorithm to transliterate each sentence. This approach ensures that complex, multi-character orthographic units (like a consonant plus its vowel diacritics) are correctly matched to their full IPA representation from the lexicon before any shorter, constituent parts are considered.

Our primary goal was to create a high-precision corpus rather than a high-recall one. To achieve this, we implemented a strict filtering step: after the pipeline was applied, we discarded any sentence that still contained orthographic characters. We identified these incomplete transliterations by checking for any remaining characters within the official Unicode ranges for Khmer (U+1780 to U+17FF) and Lao (U+0E80 to U+0EFF). This aggressive filtering strategy guarantees that every sentence in our final resource is *fully* covered by our human-annotated seed lexicon.

This process yielded a final parallel corpus of **1,314,779 sentences for Khmer** and **367,388 sentences for Lao**. While this represents a significant reduction in volume from the source corpora, it makes a crucial trade-off: we sacrifice the *quantity* of the noisy, larger datasets for the *quality* and *precision* of a fully-verified subset.

Comprehensive statistics of the resulting corpus are presented in Table 1, and sample sentence pairs are shown in Table 2. We refer to this new resource as **MekongPhon**—a clean, large-scale, and fully phonetic parallel corpus for Lao and Khmer. MekongPhon serves as a high-quality foundation for building and evaluating Grapheme-to-Phoneme (G2P) models, as demonstrated in Section 4. To facilitate transparency and future research, we release the complete pipeline, seed lexicon, and resulting corpora in our GitHub repository[4].

4. Evaluation

4.1. Evaluation Setup

To ensure a fair and rigorous comparison across models, we constructed a high-quality evaluation dataset derived from the same human-annotated seed lexicon described in Section 3.1. From each language's lexicon, we randomly held out 100 entries (representing 1.25% of Khmer and 1.66% of Lao entries) to serve as unseen data. The remaining entries were used to generate the train-

[4] https://github.com/byu-matrix-lab/MekongPhon

Language	Orthographic	IPA
Khmer	Messi ថា ក្នុង ចំណោម ខ្សែប្រយុទ្ធ ទាំងអស់ មាន តែ Ronaldo ដែល អស្ចារ្យ ជាងគេ	Messi tʰaa knoŋ cɑmnaom ksaepraɣut teaŋʔɑh mien tae Ronaldo dael ʔɑhcaa cieŋkee
Khmer	ភ្នែករបស់សត្វពស់សម្រាកនៅមុខព្រះអាទិត្យ	pnɛɛkrɔbɑhsatpʊəhsɑmraaknɨvmuk preah ʔaatit
Khmer	តើខ្ញុំនឹងមិនយល់អ្វី?	taəkɲomnɨŋmɨnyʊəlʔvəy?
Khmer	ពួកគេបរិច្ចាគបន្ទាប់ពីសាងសង់ប័ម។	puəkkeebɑɑrəccaakbɑntoappiisaaŋsɑŋ pɑɑm.
Lao	ລູກຄ້າຂອງພວກເຮົາມາຈາກຈີນແຜ່ນດິນໃຫຍ່ ແລະນອກຈາກນີ້ອາເມລິກາແລະປະເທດເອີຣົບ.	luːkˈkʰaːˈkʰɔːŋˊpʰuaːkˈhaoˀmaːˀt͡ɕaːkˌ t͡ɕiːnpʰɛːnˀdiŋŋajˀlɛˀnɔːkˈt͡ɕaːkˌniːˀ aːmeːˀliˀkaː lɛˀpaˀtʰeːtˈəːropˀ.
Lao	10 ປີແຕ່ລະຄົນ ຖ້າພວກເຂົາເຈົ້າຖືກພົບເຫັນວ່າ ມີຄວາມຜິດ.	10 piː tɛːˀlaˀkʰonˀ tʰaːˌpʰuaːkˈ kʰaoˊt͡ɕaoˈ tʰɯːkˌpʰopˀhenˊwaːˀ miːˀkʰwaːmˀpʰitˀ.
Lao	The iPhone SE ມີສອງພະລັງງານປຸງແຕ່ງແລະ ການເວລາສີ່ການປະຕິບັດຮູບພາບຂອງ iPhone 5S, ເຊັ່ນດຽວກັນກັບຊີວິດຫມໍ້ໄຟຕໍ່ໄປອີກແລ້ວ.	The iPhone SE miːˀsɔːŋˊpʰaˀlaŋˀŋaːnˀ puŋtɛːŋˀlɛˀkaːnweːˀlaːˀsiːˀkaːnpaˀtiˀbatˀ huːpˈpʰaːpˈkʰɔːŋˊ iPhone 5S, senˀdiaːw kankapˀsiːˀwitˀmɔːˌfajˀ tɔːˀpajiːkˌlɛːwˈ.
Lao	ເຕັກ ໂນ ໂລ ຊີ ການ ຮຽນ ຮູ້ ພາ ສາ ຫັນ ປ່ຽນ ຜູ້ ຮຽນ ຈາກ…	tekˀ noːˀ loːˀ siːˀ kaːn hiːanˀ huːˈ pʰaːˀ saːˊ hanˊ piaːnˀ pʰuːˌ hiːanˀ t͡ɕaːkˌ…

Table 2: Example sentences in Khmer and Lao orthography, with the corresponding IPA transliterations from the MekongPhon corpus.

Split	Khmer	Lao
Training	1,194,923	343,045
Validation	60,460	13,442
Test	60,461	13,442

Table 3: Data splits for training, validation, and testing of the G2P models.

ing corpus following the procedure described in Section 3.2. After corpus generation, we used the complete lexicon—including the held-out entries—to generate evaluation sentences. We then removed any sentence that overlapped with the training data to guarantee a fully disjoint evaluation set.

This process yielded **1,194,923** parallel training segments and **120,921** evaluation segments for Khmer, and **343,045** training segments and **26,884** evaluation segments for Lao. For each language, the evaluation set was divided evenly into validation and test subsets to ensure a balanced and representative distribution of orthographic and phonetic patterns. The final data splits are summarized in Table 3.

We use these data splits to train our sequence-to-sequence Transformer neural network and evaluate all compared systems, described in Section 4.3

4.2. Metrics

We evaluate the systems using two complementary metrics:

- **Character Error Rate (CER)** — a standard string-edit metric that measures the proportion of substitutions (S), deletions (D), and insertions (I) required to transform the predicted IPA sequence into the reference sequence, normalized by the total number of characters (N). We multiply by 100 for readability:

$$\text{CER} = \frac{S + D + I}{N} \times 100 \qquad (1)$$

- **chrF++** — a character-level F-score metric (Popović, 2017), widely used in low-resource and morphologically rich languages, as it captures fine-grained similarity beyond exact matches.

Together, these metrics assess both the strict accuracy (via CER) and the overall similarity of the predicted transcriptions (via chrF++).

4.3. Systems Compared

We compare three systems:

Epitran (Baseline). Epitran (Mortensen et al., 2018) is a widely used open-source G2P system covering over 150 languages. It serves as a natural baseline for Khmer and Lao, though its static rule mappings result in severe coverage gaps, particularly for Khmer (as discussed in Section 2.2). In our evaluation, we use the default pretrained configurations: `khm-Khmr` and `lao-Laoo`.

Rule-Based (Lao2IPA and Khm2IPA). Our rule-based transliteration tools, **Lao2IPA** and **Khm2IPA**, extend and substantially improve upon existing solutions such as Epitran by applying a series of sequential, linguistically motivated transformations.

Lao2IPA operates through modular character mappings stored in easily editable data structures, covering consonants, vowels, tone markers, and special vowel orderings. It performs transformations such as vowel–consonant reordering, tone marker repositioning, and vowel cluster merging, ultimately producing syllable-level representations that capture the complex tonal and segmental structure of Lao. Each syllable is represented by a dedicated `Syllable` class, which encodes relevant phonological information—including consonant class, vowel length, and dead-syllable status—to generate accurate tone realizations according to the Vientiane dialect. The system optionally allows tone inclusion or omission, making it flexible for phonetic versus phonemic transcription needs. Lao2IPA is resilient to noisy input, performing cleaning steps (e.g., merging mis-entered vowel characters) and fallback mappings for malformed clusters.

For Khmer, we developed a parallel system, **Khm2IPA**, which adapts the core architecture of Lao2IPA but omits tone modeling, as Khmer is non-tonal. Instead, it incorporates the language's dual vowel series—corresponding to the inherent vowel sounds associated with each consonant series (A-series vs. O-series). Our implementation adds conditional mappings for both vowel types, applying contextual rules that determine the appropriate vowel pronunciation based on consonant strength (strong vs. weak) and orthographic context. Together, Lao2IPA and Khm2IPA constitute robust, extensible transliteration frameworks capable of fully converting Lao and Khmer orthography into standardized IPA representations, offering higher coverage and linguistic accuracy than prior rule-based approaches.

Neural Model. Our neural G2P system is a sequence-to-sequence Transformer implemented with the HuggingFace `EncoderDecoderModel`. We initialize a BERT-style encoder and a BERT-LM decoder from configs and train *from scratch* (no pretrained weights), using language-specific tokenizers for source orthography and target IPA. Datasets are tokenized with separate **ByteBPE** models (vocab = 1,000 each) for the encoder (orthography) and decoder (IPA). The training loop uses cross-entropy with teacher forcing; validation loss is monitored to select the best checkpoint.

We scale the model to a **6-layer** encoder and decoder with **8** attention heads, **hidden size 512**, and **intermediate size 1024**; max length is 128. Training is launched with `accelerate` to convergence using 4 H100 GPUs, using a learning rate of **1e-4**, **batch size 128** and the script's evaluation cadence of every **2,000** steps, with early stopping (patience = 3) and best-model selection on validation loss. At inference, we decode with **beam search (beam=5)** over the decoder tokenizer. Implementation details (model construction, training arguments, and callbacks, including beam-aware generation during eval) follow our training file.

Three models are trained: a Khmer G2P model, a Lao G2P model with tones, and a Lao G2P model without tones, allowing us to examine the influence of tone information on transliteration performance and have a fair comparison to Epitran.

5. Results and Discussion

Using the held-out test sets described in Section 4, we evaluate three systems: the baseline **Epitran**, our linguistically informed **rule-based systems**, and the **Transformer models** trained on the MekongPhon orthographic–IPA data. Results are summarized in Table 4, including inference speed measured in sentences per second on a single CPU.

5.1. Overall Performance

Epitran performs poorly across all languages, with CER values exceeding 106 for Khmer and above 33 for Lao (with tones), and chrF++ scores below 30 in both cases, reflecting highly inaccurate phonetic transcriptions. For Khmer specifically, Epitran's failures are systematic: it retains orthographic vowel signs in the output while simultaneously producing inherent vowels for each consonant, yielding transcriptions such as ចិនទាំង → cɑːិnɔːtɔːាŋɔː that are phonetically incoherent. This reveals a fundamental incompatibility between Epitran's character-level mapping approach and the abugida structure of Khmer, where consonant–vowel interactions must be resolved jointly rather than sequentially. This confirms that the static map-based approach of Epitran is insufficient for complex writing systems such as Khmer

	Khmer			Lao (tones)			Lao (no tones)		
Model	**CER ↓**	**chrF++ ↑**	**Sent/s**	**CER ↓**	**chrF++ ↑**	**Sent/s**	**CER ↓**	**chrF++ ↑**	**Sent/s**
Epitran	106.57	7.82	2076.5	33.48	28.36	2525.8	12.45	67.66	2570.4
Rule-Based	35.49	32.15	619.5	9.64	76.31	5191.5	5.96	83.11	16608.6
Neural	1.42	98.24	2.0	1.39	95.74	1.6	1.39	96.08	2.5

Table 4: Evaluation results for Khmer and Lao G2P systems. Lower CER indicates better phonetic accuracy; higher chrF++ indicates greater character-level similarity. Sentences per second (Sent/s) inference speed are calculated on the same 1000 sentences for each language using the exact same CPU.

and Lao, which rely on positional vowel signs, consonant series, and tonal interactions.

The comparatively higher error rate of the Khmer rule-based system can be attributed to a systematic mismatch in IPA typology: Khm2IPA is derived from the Wikipedia IPA/Khmer mapping conventions,[5] which differ in numerous small but consequential ways from the transcription conventions used in the 1977 Cambodian-English dictionary (Headley, 1977) that underlies our test set. As a result, the system may produce phonetically reasonable transcriptions that nonetheless diverge from the reference in consistent, predictable ways. We note that both systems may be appropriate depending on the IPA convention desired by the end user; researchers whose work aligns with the Wikipedia conventions may find the rule-based system preferable, particularly given its substantially higher inference speed.

5.2. Neural Models

The Transformer models trained on the MekongPhon corpus outperform both baselines by a large margin, achieving near-perfect transcriptions with CER below 2% and chrF++ above 95 for all three languages. These results indicate that the MekongPhon data provide sufficient coverage and consistency for data-driven G2P learning, even in traditionally low-resource languages. Interestingly, the model performs similarly across the tonal and non-tonal Lao variants, suggesting that tone prediction can be learned effectively when provided with explicit tone markers in the target IPA.

5.3. Efficiency Trade-offs

While the neural models deliver state-of-the-art accuracy, this comes at a significant computational cost. On CPU inference, the Transformer models process only 1–3 sentences per second, compared to thousands for the rule-based systems and Epitran. GPU acceleration substantially narrows this gap: on a single A100 GPU with a batch size of 256, inference speed increases to roughly 78–115 sentences per second. Nonetheless, the difference in throughput underscores a practical trade-off between computational efficiency and accuracy. For real-time or on-device applications, our rule-based systems offer a strong balance between performance and resource efficiency, while the neural models are better suited to high-accuracy offline processing or integration into downstream models.

5.4. Qualitative Analysis

To complement the quantitative results, we conducted a qualitative analysis of model errors for both Khmer and Lao, examining the ten sentences with the highest CER for each system on the held-out test set.

Lao. Error analysis for Epitran shows a consistent omission of tones in the output, which is the primary cause of increased CER for the test set that includes tones. Across both tonal and non-tonal test sets, Epitran also struggled with compound consonants and vowels, such as in the words ເຫລືອງ and ຫວັ່ງ. Additionally, Epitran failed to correctly swap the preorder ໃ vowel.

Both Epitran and our rule-based model struggled to differentiate between instances where the ອ character was used as a vowel or a consonant. Both also experienced higher CER scores due to a mismatch in the phonetic transcription systems of the models and the test set, as described in Section 5.1.

In addition to this difficulty resolving the consonant-vowel ambiguity for characters such as ອ and ອ, the rule-based model struggled with incorrectly combining consonants into a cluster and not being able to adapt to exceptions for tone calculations, such as words borrowed from English and names.

The neural model, in comparison, struggled mainly with hallucination and truncated outputs, especially when given words that were not found in the training data. We also observed many instances where the model incorrectly separated syllables, placing tones in the middle of syllables and sometimes hallucinating vowels or conso-

[5] https://en.wikipedia.org/wiki/Help:IPA/Khmer

nants. Because it was trained on data in the same format as the test sets, the neural model rarely experienced issues with mismatches between the phonetic transcription systems of the model and the test set. The downside of this is the model's tendency to overfit to the training data, even when the training data is incorrect.

All three models struggled with rare vowel combinations that were not seen in design and training. In this situation, the two rule-based models would transliterate each vowel character separately, while the neural model would begin hallucinating the remainder of the output.

Khmer. Error analysis for Epitran confirms the systematic failures described in Section 5.1, with the character-level mapping approach consistently producing phonetically incoherent output due to its inability to resolve abugida-level consonant–vowel interactions.

For the neural model, as with the Lao model, the majority of errors occur in longer sequences, likely reflecting a distributional mismatch in which the training data underrepresents extended utterances, causing the model to produce truncated or incomplete transcriptions. Occasional errors also involve the incorrect addition of vowel sounds after word-final consonants that should remain unreleased or silent in Khmer phonology.

Errors in the rule-based system are largely attributable to systematic divergences in IPA convention rather than phonological misanalysis: as noted in Section 5.1, the mappings derived from the Wikipedia IPA/Khmer scheme differ in consistent, predictable ways from those of the Headley reference transcriptions, yielding outputs that are phonetically defensible but metrically penalized.

5.5. Implications

These findings showcase the strength of the MekongPhon corpus as a high-quality training resource. The exceptionally low CER and high chrF++ achieved by the neural systems demonstrate that a well-curated, phonemically consistent dataset can enable accurate G2P modeling for previously underserved languages.

While the MekongPhon corpus is constructed by filtering to sentences fully covered by the seed lexicon, this does not mean the lexicon alone is sufficient for general-purpose G2P. The corpus expansion pipeline is designed for high-precision data generation, not inference: it deliberately discards any sentence containing characters outside the seed lexicon's coverage, which means the resulting training data is clean but the underlying trie is not a deployable G2P system for arbitrary input. In practice, real-world text inevitably contains words absent from the lexicon, such as neologisms, loanwords, named entities, and non-standard spellings. The corpus expansion and trie-based algorithm cannot handle and would simply fail on such examples. A Transformer model trained on MekongPhon, by contrast, learns productive phonological patterns from the data and can generalize to unseen forms. The MekongPhon corpus thus serves a dual purpose: it is both a high-quality released resource and the training foundation for a more robust and general-purpose G2P system than the lexicon alone could support.

6. Conclusion

In this work, we introduced MekongPhon, the first large-scale, high-quality parallel IPA corpus for Khmer and Lao, addressing the critical lack of phonetic resources for these low-resource languages. Using this corpus, we developed both rule-based and neural G2P systems that dramatically outperform existing baselines such as Epitran, achieving near-perfect transliteration accuracy on a strong test set.

Beyond Khmer and Lao, our methodology is general and can be applied to any language that has a lexicon of word–IPA pairs and a sufficiently large monolingual corpus. By combining a verified seed lexicon, efficient trie-based expansion, and filtered corpus generation, it is possible to create high-quality phonetic parallel data for other low-resource languages. Such corpora can in turn be used to train language-specific or multilingual neural transliteration models, facilitating broader phonetic modeling across underrepresented scripts.

While our neural models achieve exceptional accuracy, they remain computationally intensive for real-time use. Future work will explore lightweight Transformer variants, knowledge distillation, or multilingual knowledge-transfer approaches to reduce inference cost. Additionally, Large Language Models (LLMs) could take advantage of this domain specific data for various NLP and multi-modal tasks.

We hope that researchers will leverage this resource to advance speech and language technologies for Lao and Khmer, enabling further progress in low-resource language research.

7. Limitations

Although MekongPhon substantially improves IPA resources for Lao and Khmer, several limitations remain. Our web-scraping approach to the SEALang website was conservative, leaving a significant portion of high-quality lexical and phonological information unused. Both the rule-based and neural transliteration systems could also be

enhanced: the rule-based scripts prioritize efficiency at the cost of phonetic accuracy, while the neural models, despite strong performance, would benefit from architectural tuning and additional high-quality data to better handle rare or irregular orthographic patterns. The neural models are also sensitive to incorrect mappings in the lexicon, and may produce incorrect transliterations if inaccurate lexicon data is present. By nature, both the rule-based and neural systems are also generally unable to produce the surface representation of words, since they operate based on documented phonology. The tones used in Lao are from the Vientiane dialect and may differ from other dialects or time periods.

In addition, the OPUS datasets used for training—particularly NLLB and ParaCrawl Bonus—contain considerable noise and alignment inconsistencies. This data quality issue likely constrains the upper bound of our model performance. Some lexicon mappings contain atypical IPA representations for vowels, leading to higher error rates. Future work will focus on more thorough data cleaning, expanding lexical coverage, and exploring improved G2P modeling techniques to further advance phonetic resource quality for low-resource languages in the Mekong region.

8. Bibliographical References

Shiyang Cheng, Pengcheng Zhu, Jueting Liu, and Zehua Wang. 2024. A survey of grapheme-to-phoneme conversion methods. *Applied Sciences*, 14(24):11790.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Robert Kirk Headley. 1977. *Cambodian-English dictionary*. Catholic University Press.

International Phonetic Association. 2020. The international phonetic alphabet (ipa) chart. Online resource. Accessed: 2025-10-23.

Allen D Kerr. 1972. *Lao-English dictionary*. Catholic University Press.

Philipp Koehn. 2024. Neural methods for aligning large-scale parallel corpora from the web for south and East Asian languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1454–1466, Miami, Florida, USA. Association for Computational Linguistics.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Nankai Lin, Yingwen Fu, Chuwei Chen, Ziyu Yang, and Shengyi Jiang. 2022. LaoPLM: Pre-trained language models for Lao. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6506–6512, Marseille, France. European Language Resources Association.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

David R Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision g2p for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2020. Ccmatrix: Mining billions of high-quality parallel sentences on the web.

Kak Soky, Xugang Lu, Peng Shen, Hiroaki Kato, Hisashi Kawai, Chuon Vanna, and Vichet

Chea. 2016. Building wfst based grapheme to phoneme conversion for khmer. *Proc. Khmer Natural Language Processing (KNLP)*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).