Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Dialect Identification from Automatically Transcribed Speech

Kellen Parker van Dam
Chair for Multilingual Computational Linguistics
University of Passau

This study investigates whether automatic speech recognition (ASR) output can support rapid dialect identification in Khiamniungan, an under-documented Tibeto-Burman language spoken across Northeast India and Myanmar. Using transcripts produced by an existing wav2vec 2.0 ASR model, we train two text-based dialect identification systems on 83.93 minutes of speech from six speakers representing three dialects: Thang, Nokhu, and Peshu. The systems use either regularized logistic regression over syllable, syllable-pair, and phoneme n-gram features, or character n-gram language models scored by perplexity. Under file-grouped cross-validation, both models correctly classify 37 of 38 held-out files, achieving 0.974 accuracy and 0.975 macro F1. The models also recover the same dialect relationships, grouping Thang and Nokhu against Peshu. On two previously unseen broadcasts, including one by a speaker whose home variety differs from the variety spoken on air, both systems correctly identify Thang from approximately half a minute of automatic transcription. These results show that ASR transcripts can provide sufficient phonological and lexical information for efficient dialect identification without additional neural modeling, offering a lightweight tool for organizing and comparing documentary language archives.

1 Introduction

This study investigates how well rapid dialect identification of audio files based on a trained ASR model for the Khiamniungan language works (Glottolog: khia1236), using a logistic-regression classifier over morpheme, morpheme-bigram and phoneme n-gram features, and a per-dialect character n-gram language model scored by perplexity.

Computer-Assisted Language Comparison in Practice
Volume 9, Number 1
URL: https://calc.hypotheses.org/9593 DOI: 10.15475/calcip.2026.2.3

Published on 29/04/2026

The transcripts this study works from are the output of the wav2vec 2.0 model of Thaam et al. (2026) with no additional fine-tuning. That model was built by a staged procedure for multi-dialectal Khiamniungan across the Khiamniungan speaking area of Northeast India and Myanmar. This study builds on that by training dialect detection models on transcripts from six news reporters' broadcasts. The broadcasts are taken from *AIR Akashvani Radio Kohima* youtube.com/c/AIRNEWSKOHIMA, a public radio broadcaster in the state of Nagaland, India, which presents dialectal news reports for a number of indigenous languages.

The Khiamniungan broadcasts vary in dialect from speaker to speaker. We take three dialects as representative of the three Khiamniungan sub-branches (van Dam 2026). These are Thang (Northern branch), Nokhu (Eastern branch) and Peshu (Southern branch). Additional named varieties exist within the three branches, but for our purposes here, we deal only with these three dialects for the initial study. We rely only on news broadcasts here, enforcing a consistent genre and so a consistent lexicon.

code	**dialect**	**branch**	**sex**	**village**	**duration**	**syllables**
NM	Thang	Northern	m	Thang (Noklak)	16.51	4,703
NF	Thang	Northern	f	Thang (Noklak)	11.50	3,421
EM	Nokhu	Eastern	m	Nokhu	14.52	3,720
EF	Nokhu	Eastern	f	Nokhu	15.95	4,107
SM	Peshu	Southern	m	Tsoiphu-Kingjung	15.14	3,974
SF	Peshu	Southern	f	Peshu	10.31	2,748
				total	**83.93**	**22,673**

Table 1: Training inventory

Two reporters were chosen for each dialect, one man and one woman, for six speakers in total. Table 1 lists them by code, with the amount of speech each contributes. The dialect assigned to each is the variety of the village the reporter is known to come from, as set out in §3.1.

The three dialects are far enough apart segmentally that the frequency of syllable shapes and of adjacent syllable pairs should follow dialect without any help from vocabulary. Table 2 gives five cognate sets across Thang, the neighboring northern variety Wolam, Nokhu, Peshu, and the neighboring Pathso variety. Even when the etymon is shared, the syllabic reflexes differ in onset, rhyme and at times in syllable count. A model over syllable n-gram frequencies has this segmental spread to draw on before any dialect-specific word is counted.

gloss	Thang	Wolam	Nokhu	Peshu	Pathso
‘day’	*ʔa.ʧe.kʰaiʉ*	*ʔa.te.kʰai*	*ʔa.tsə.kʰeʊ*	*ʔɔ.tsi.kʰai*	*ʔa.tsi.kʰiʊ*
‘person’	*kʰaiu.ɲaʔ*	*kʰai.ɲaʔ*	*kʰeʊ.ɲaʔ*	*kə.ɲɔʔ*	*kʰiʊ.ɲɔʔ*
‘one’s own’	*mə.ʃjn*	*mə.hin*	*me.ʃin*	*mə.sa*	*mə.sa.nɔ*
‘about’	*tʰaiʉ.tʰaʔ*	*tʰai.tʰaʔ*	*tʰeʊ tʰaʔ*	*tʰai.tʰɔʔ*	*tʰiʊ.tʰɔʔ*
‘information’	*ŋiʊ.ŋem*	*ŋiʊ.ŋim*	*ŋiʊ.ŋem*	*ŋe.ŋam*	*ŋiʊ.ŋem*
‘five’	*pə.ŋɔ*	*pə.ŋɔ*	*mə.ŋɔ*	*mə.ŋɔ*	*mə.ŋɔ*

Table 2: Phonological comparison across dialects

2 Related Work

Identifying a variety from a stretch of text is a long-standing task, surveyed by Jauhiainen et al. (2019) and run as a shared task across the VarDial evaluation campaigns (Zampieri et al. 2017). The closest of those tasks to the present study is German Dialect Identification, where systems label transcripts of spoken Swiss German as Basel, Bern, Lucerne or Zurich. Good results are attainable via a meta-classifier over character and word n-grams (Malmasi & Zampieri 2017).

Our speech pipeline used here has an old precedent. Zissman (1996) found that phone recognition followed by language modeling (PRLM), i.e. running a phone recogniser over an utterance and classifying the resulting symbol string with n-gram language models, outperformed direct acoustic classification for language identification, with parallel PRLM the strongest of the four approaches compared. Current systems classify from acoustic embeddings, with x-vectors the standard representation (Snyder et al. 2018). Ali et al. (2016) identify five Arabic dialects in broadcast speech by combining phonetic and lexical features drawn from an ASR system with i-vector features.

Aggregate dialectometry measures distance between varieties instead of labelling recordings. Nerbonne & Heeringa (1997) computed Levenshtein distances over phonetic transcriptions of parallel word lists (see also Wieling & Nerbonne 2015). Those methods need comparable items elicited at every site. This would be possible for a Khiamniungan dialectometric study, but the goal here is to place as-yet-unseen recordings within the Khiamniungan family tree.

Automatic transcription is also becoming more and more established in documentation and description (Michaud et al. 2018; Adams et al. 2018). Self-supervised multilingual pretraining has since cut the amount of labelled audio a new language needs (Conneau et al. 2021), and the model that produced the transcripts used here is one of that kind, trained by Thaam et al. (2026) and applied here unchanged.

To our knowledge, these previous lines of work have not been joined for a language in Khiamniungan’s position. Text dialect identification presupposes a written corpus, and

speech dialect identification presupposes enough labelled audio per variety to train an acoustic model for each. What we test is whether the phonemic output of a documentary ASR model — the same model and the same output a documenter already produces — carries enough signal to sort recordings by variety. If it does, dialect identification costs nothing beyond the transcription step for any language that has reached the point of having a working ASR model, and that set is growing. We found no earlier study closing the loop this way, though not finding one is not proof that none exists.

3 Methodology

3.1 Data

From each of these three dialects we take broadcasts by two speakers, one male and one female. Speakers' dialects were identified by a native speaker of Khiamniungan with formal linguistic training and expertise in Khiamniungan dialect comparison, to whom the reporters are known personally. As such, the label is not a perceptual judgment made from the recording but a known fact of each speaker. The broadcasts were then checked by the same speaker to confirm that the dialect spoken by the presenter was their own. This is true for all six speakers in the training data, but not the testing data described in later sections.

For each speaker we aim at roughly 15 minutes of speech. The material carrying a dialect annotation comes to 83.93 minutes and 22,673 syllables, split across the six speakers as given in Table 1. The speech is transcribed phonemically, omitting code-mixing and personal names. Proper nouns that fit within the Khiamniungan phonotactic restrictions (such as the name of the city Kohima) are retained in the transcription.

While fifteen minutes is the target, the syllable densities differ from speaker to speaker. Before any model is fitted we therefore cap the corpus, reducing each dialect to the syllable-token total of the smallest and dividing that allowance evenly between its two speakers. The dialect totals are 7,827 syllables for Nokhu, 8,124 for Thang and 6,722 for Peshu, so the cap sits at 6,722 and discards 2,449 syllables, 10.8% of the corpus. A speaker over quota has each of their files thinned at the same rate, keeping intervals spaced along the whole recording, so no file drops out and no stretch of a broadcast is preferred over another; because the thinning works file by file, a dialect comes to rest a little above the cap and not exactly on it. The cap is applied before the vectoriser is fitted, so the feature vocabulary, the per-dialect language models and the marker test of §3.4 all see a corpus in which no dialect can be separated on the strength of how much of it was transcribed.

To further ensure that we are comparing like with like, transcriptions were normalized to the Cross-Linguistic Transcription Systems standard, CLTS (Anderson et al. 2018). We then automatically syllabify the transcript based on the maximum onset principle and sonority hierarchy (Clements 1990), within the syllable canon described by Thaam & Kevichüsa-Ezung (2024). Syllabification is then checked by a human. This is done in a plain text file, since the only focus of the model is the syllable transcription. As is common across Sino-Tibetan, a Khiamniungan syllable is also typically a single morpheme (but cf. van Dam & Thaam 2026a), so a model over syllables is at the same time a model over morphemes. These transcripts come out of the same transcription effort as the ASR corpus of §1, so transcription policy and encoding are the same on both sides of the pipeline and the dialect models are fitted on text of the kind the ASR model emits.

3.2 Features

From the syllabified texts we train two models on the IPA transcript alone, both predicting the dialect as named by the file. The idea is that collocation patterns will show dialectal signal, partly through phonology and partly through lexical differences, though in the end the two amount to the same thing: frequency of syllable shapes in adjacent contexts as they occur in the three dialects.

Each interval of the syllabified transcript is one training instance. Its label is read from a parallel dialect tier: an interval takes the dialect whose annotation covers its midpoint, so a file in which the variety shifts partway through is labelled interval by interval and not as a whole.

Three feature blocks are extracted and concatenated: TF-IDF weights over single syllables, over adjacent syllable pairs, and over character n-grams of up to three symbols taken within syllable boundaries, so a shared onset or rhyme is still visible when the whole syllable is unattested elsewhere in the training data (Salton & Buckley 1988). All three blocks use sublinear term frequency, which log-scales repetition so that a syllable said twenty times in one interval does not count as twenty independent observations. The syllable-pair block admits a pair only once it has appeared in two separate intervals.

The result tables of §4 label these three blocks morpheme, morpheme pair and phoneme n-gram, on the syllable-morpheme identity set out in §3.1. Reduplication is productive in Khiamniungan across several formation types (Thaam 2025), so a good part of what the syllable-pair block counts is reduplicative. The etyma behind those forms are shared across the dialects while their phonological reflexes are not, which makes them a likely source of signal for this task. §4.1 reports each of the three blocks cross-validated on its own alongside the union, so what each contributes can be read off.

3.3 Classifiers and Validation

The first model is a regularized logistic regression over these sparse features, predicting the dialect. Class weights are set inversely to each dialect's frequency among the training rows, which is a second correction on top of the token cap of §3.1: the cap equalizes the dialects by syllable count, the class weights equalize them by interval count, and the two come apart because interval length varies with speaker and with genre.

The second model is a set of per-dialect character language models, each of order 5 with Witten-Bell interpolation (Witten & Bell 1991). A recording is scored against every dialect's model by perplexity and assigned to the one it fits best; the report gives that perplexity and its margin over the runner-up, with no probability attached to either. Perplexities are inverted and normalized into a probability vector at one point only, inside cross-validation, where that vector supplies the distances for the language model's dendrogram in §3.5. It reads the transcript as an unbroken character stream whereas the classifier reads syllables and syllable pairs, so the two can disagree.

Validation uses GroupKFold with five folds and the source file as the group (Pedregosa et al. 2011). For the classifier, accuracy is reported both per interval and after pooling all intervals of a file into a single document; the language models are scored per whole file. Grouping by file keeps whole broadcasts intact across the split, with both speakers of a dialect represented across the folds.

From the classifier's validated probabilities, pooled to one document per file, we also fix a confidence threshold, and any later estimate falling below it is flagged as uncertain. It is the lowest cut-off on a 0.30 to 0.95 grid whose retained guesses are at least 85% accurate. On this corpus that is 0.30, the floor of the grid.

3.4 Significant Markers

The markers reported alongside the estimate are not the regression coefficients. They are found in a separate pass, using a one-sided Fisher exact test for each syllable against each class (Agresti 2001), Benjamini-Hochberg control of the false discovery rate at $q = 0.05$ (Benjamini & Hochberg 1995), and a weighted log-odds effect size (Monroe et al. 2008); a marker must clear both the $q = 0.05$ cutoff and a weighted log-odds z of 2.0, so any marker printed in the report is one whose association with a dialect holds up under testing on its own. The counting unit in this pass is the file, not the interval: a syllable is counted once per broadcast it occurs in. Counting intervals instead would treat the repetitions within one speaker's broadcast as independent observations and return far more markers than the data support.

3.5 The Dialect Tree

The long-term goal is a more fine-grained study with recordings from each of the major villages and each of the major dialect groups, sub-varieties included, so the cross-validated probabilities are also reused to build a dendrogram of the dialects, one from each model. Distance between two dialects comes from their confusability, the probability mass a model assigns to each when the other is the true label, normalized against the mass it keeps on the correct label, and the resulting matrix is clustered by UPGMA (Sokal & Michener 1958). Clade support is estimated by resampling the source files two hundred times and recording how often each clade reappears (Felsenstein 1985). What this measures is which dialects a model confuses on this corpus, so it is not an independent test of the subgrouping assumed in §1.

3.6 Applying the Model

To identify the dialect of a new recording, the audio is passed through the ASR model of §1 and both dialect models are applied to the pooled transcript it returns. The recordings used for this were never part of any training data, for the ASR model or for the dialect models, and their dialect is already known from the reporter, so each estimate can be scored against a label neither stage was ever given.

The run writes a markdown report giving each model’s estimate and noting whether the two agree, with the classifier’s probability flagged when it falls below the threshold fixed during validation. The report also names the features and the significant markers behind each estimate and scores every interval on its own, so the evidence can be inspected where it falls in the recording.

Both models are forced choice: a recording in a variety absent from the training data will still be assigned to one of the trained dialects, and the broadcast stream does contain such varieties. The confidence threshold gives a way to abstain, but it was calibrated on held-out files whose dialects were all present in training, so it is a check on low-confidence decisions within the known set and not a test for varieties outside it. We do not test that case here, for the reason given in §1.

3.7 Implementation

The pipeline is implemented in Python as a standalone script, available at codeberg.org/besra/supplementary_code. Feature extraction, classification and cross-validation use scikit-learn (Pedregosa et al. 2011): TfidfVectorizer builds the three feature blocks of §3.2, FeatureUnion concatenates them, LogisticRegression fits the classifier, and GroupKFold supplies the file-grouped folds used throughout §3.3–3.5. Numeric arrays and the vectorised operations over them are handled by NumPy (Harris

et al. 2020). SciPy (Virtanen et al. 2020) supplies the Fisher exact test and Benjamini-Hochberg correction of §3.4, and the UPGMA linkage and distance-matrix routines behind the dendrograms of §3.5. Trained models, including the vectoriser, the classifier, the per-dialect language models, the marker list, the confidence threshold and both trees, are serialised to a single file with joblib, which is also what lets the identification step in §3.6 load a model fitted once and reuse it across recordings without retraining. Dendrogram and dialect-probability figures are rendered with Matplotlib (Hunter 2007). The plugin's dialog and popups are built on PySide6, the Python bindings for Qt, which is BESRA's own GUI toolkit rather than a component of the dialect model as such.

4 Results

4.1 Cross-Validated Accuracy

The corpus holds 3,412 labeled intervals across 38 files. The token cap of §3.1 sets the ceiling at 6,722 morpheme tokens, the total for Peshu dialect, and leaves the three dialects at 6,760 tokens for Nokhu, 6,742 for Thang and 6,722 for Peshu. Table 3 gives both models against the majority-class baseline under GroupKFold with the source file as the group, and Table 4 breaks the classifier down by class per interval.

model	unit	n	macro F1	accuracy	majority baseline
logistic regression	interval	3,412	0.791	0.788	0.409
logistic regression	whole file	38	0.975	0.974	0.395
character LM, order 5	whole file	38	0.975	0.974	0.395

Table 3: Model scores against the majority-class baseline

class	precision	recall	F1	support
Nokhu	0.765	0.779	0.772	1,159
Peshu	0.798	0.849	0.822	859
Thang	0.801	0.757	0.778	1,394

Table 4: Precision and recall for dialect models

Per interval, the classifier reaches 0.791 macro F1 against a 0.409 baseline, and the three dialects behave alike, with recall between 0.757 and 0.849. The confusion of Table 5 is not symmetric in weight. Thang and Nokhu trade the bulk of their errors with each other, while Peshu is the dialect least often confused in either direction, holding the highest recall and F1 of the three and drawing the fewest errors from the other two. This supports earlier descriptions of Khiamniungan dialects as being split primarily north/south, with Eastern (Nokhu) as a transitional branch holding features of the other two, although in this case more strongly those of the Northern branch.

predicted → true ↓	**Nokhu**	**Peshu**	**Thang**
Nokhu	903	66	190
Peshu	58	729	72
Thang	220	119	1055

Table 5: Cross-dialect confusion matrix

The classifier labels held-out intervals from 37 of the 38 files correctly, at 0.974 accuracy and 0.975 macro F1, and the single error is a Nokhu file taken for Peshu. The language models reach the same 0.974 on the same files and make the same single error, so at file level the two models are not distinguishable on this corpus. Cross-perplexity, in Table 6, separates the dialects by a wide margin, with self-perplexity between 3.0 and 3.5 and every off-diagonal cell between 9.6 and 15.0.

model text ↓	→ Nokhu	Peshu	Thang
Nokhu	3.4	15.0	12.5
Peshu	10.4	3.0	9.6
Thang	12.3	14.0	3.5

Table 6: Dialectal cross-perplexity scores

Each of the three feature blocks was then cross-validated on its own, on the same folds and with everything else held as it was. No single block accounts for the per-interval result. Character n-grams taken within syllable boundaries do a little better on their own than whole syllables. The syllable-morpheme identity motivates the syllable block, and what separates these dialects is as much sub-syllabic as it is a question of which syllables occur. Syllable pairs are the weakest block and still reach 0.658, so the fixed and reduplicative sequences described in §3.2 carry dialect information on their own.

4.2 Significant Markers

The marker test of §3.4 returns 51 syllables whose distribution over files separates a dialect, twenty-four of them occurring only in the dialect they point to. Table 7 gives the four strongest for each dialect by weighted log-odds z, with the number of files of that dialect the syllable occurs in, the number of files elsewhere, and its exclusivity, the share of its tokens falling in the dialect it points to. The full list of 51 is in the training report the pipeline writes and is not reproduced here.

morph	points to	files in class	files elsewhere	z	exclusivity
tsən	Thang	9	2	10.46	0.985
san	Thang	9	0	5.86	1.000
jam	Thang	8	0	4.92	1.000
tʰaiʊ	Thang	8	1	4.82	0.977
tʃə	Peshu	10	6	7.86	0.938
hin	Peshu	8	2	6.17	0.949
tʰɔ	Peshu	7	2	4.60	0.953
aʊ	Peshu	10	8	4.43	0.644
neʊ	Nokhu	9	0	6.58	1.000
tʰeʊ	Nokhu	9	2	4.71	0.956
niʊ	Nokhu	8	2	4.27	0.947
teʊ	Nokhu	10	3	4.10	0.921

Table 7: Candidate weights by morpheme

Nokhu yields the fewest markers at ten and also has the lowest per-interval F1 at 0.772, but Thang yields the most at twenty-four from the fewest files while sitting in the middle at 0.778, and Peshu is the most separable of the three at 0.822 on seventeen. The two counts measure different things. What the test returns is a list of syllables whose distribution over whole files is significant, and a syllable can carry a great deal of the dialect signal in the classifier without appearing here at all, since the classifier sees syllable pairs and character n-grams as well.

4.3 The Dialect Tree

Both dendrograms have the same topology, with Peshu outside a Nokhu-Thang node. Support on that internal node is 100% from the classifier and 89% from the language model over 200 resamples of the files; the root is not given a support figure, since with three taxa it is recovered in every replicate by construction. In Newick, with node heights as UPGMA merge distances:

```
(peshu:0.4082,(nokhu:0.3374,thang:0.3374)100:0.0708)100;   classifier
(peshu:0.2685,(nokhu:0.2443,thang:0.2443)89:0.0242)100;    language model
```

Nokhu and Thang are the nearest pair of the three under both models. The classifier separates nearest from furthest by 0.158 in confusability distance and the language model by 0.050, and that flatter spread is where the language model’s lower internal support comes from, though at 89% it still recovers the clade in nearly nine replicates out of ten.

The topology places the Northern and Eastern varieties together against the Southern one, which is the shape the subgrouping of §1 would predict, but the agreement is not

evidence for the subgrouping: the distances measure which dialects these two models confuse on this corpus, and with three dialects what the two trees agree on amounts to a single binary split.

4.4 Unseen Data

Two broadcasts were passed through the pipeline of §3.6, from audio through the ASR model to the two dialect models, with no hand correction at any point. Neither the recordings nor the speakers have ever been in any training data: both sit outside the corpus of Thaam et al. (2026) and outside the six-speaker corpus of §3.1, so both verdicts below are given on a voice the models have never heard. Both are Thang broadcasts and the two speakers stand in different relations to the dialect.

The reporter in 5CECu8KCMas is from Thang and broadcasts in Thang. Of greater interest for this approach is that the reporter in 0eZNh0krlVc is from Peshu but with considerable Nokhu dialectal contact, and broadcasting in Thang. The Thang scores for this broadcaster are less strongly certain about Thang being the dialect, indicating the presence of a more southern accent in her broadcast as confirmed by a native dialectal expert. Taken together these scores ask whether the models label the variety on the recording or the speaker behind it. Both models label both recordings Thang and agree with each other on both.

The first recording runs 615.02 seconds and holds 23 intervals, of which 11 carry Khiamniungan transcript and were scored, 116 syllables over 24.1 seconds of speech. The second runs 591.04 seconds and holds 22 intervals, again 11 of them scored, 101 syllables over 24.4 seconds. Table 8 gives both models over all three dialects for each recording, identified by their unique YouTube identifiers 5CECu8KCMas and 0eZNh0krlVc: the classifier’s mean probability across the scored intervals and the number of intervals each dialect wins outright, then the language model’s perplexity over the pooled transcript.

recording	dialect	mean interval probability	intervals won	perplexity
5CECu8KCMas	Thang	0.730	10	5.91
	Nokhu	0.228	1	9.97
	Peshu	0.042	0	15.82
0eZNh0krlVc	Thang	0.624	8	6.51
	Nokhu	0.270	3	8.64
	Peshu	0.106	0	11.23

Table 8: Scores for recordings 5CECu8KCMas and 0eZNh0krlVc

Pooled over the whole transcript the classifier returns Thang at p = 0.974 for the first recording and p = 0.831 for the second, both above the 0.30 threshold and so unflagged.

The Thang evidence is weaker in the second recording, where the features present sum to +0.778 toward Thang against +2.108 in the first, and where Peshu's features outweigh Thang's until the fitted intercepts are added. The largest single contribution in both recordings is the morpheme tsən, and pə.ŋɔ is the largest syllable pair in both; hi and the phoneme n-gram ␣h argue hardest against Thang in both.

Eight of the 51 markers of §4.2 occur in the first recording, in fifteen tokens over seven intervals, seven of them pointing to Thang and one to Nokhu. Nine occur in the second, in twelve tokens over seven intervals, six pointing to Thang, two to Peshu and one to Nokhu. Table 9 gives the union of the two lists, with the weighted log-odds z as fitted in training and the token count from each recording; every marker here occurs in as many intervals as it has tokens.

morpheme	points to	tokens in 5CECu8KCMas	tokens in 0eZNh0krlVc	z
tsən	Thang	4	2	10.46
san	Thang	0	1	5.86
jam	Thang	1	0	4.92
tʰaiʊ	Thang	0	1	4.82
tʰɔ	Peshu	0	1	4.60
aʊ	Peshu	0	1	4.43
iʊ	Thang	1	0	3.87
kiʊ	Thang	3	2	3.73
ŋem	Nokhu	3	2	3.40
teŋ	Thang	0	1	3.36
taiʊ	Thang	1	0	3.26
tsʰai	Thang	1	0	2.37
ɲat	Thang	1	1	2.12

Table 9: Weighted log-odds z and token count from each recording

The per-feature accounting behind each verdict and the score for every interval taken on its own are in the report the pipeline writes for each recording and are not reproduced here. The pooled verdicts above are taken over all the text at once and are not votes over the intervals.

For the second speaker the pair answers the question it was chosen to ask. Peshu is the variety the reporter grew up speaking, and it takes the smallest share of the classifier's probability mass at 0.106, wins no interval and sits last on perplexity. Two of the twelve marker tokens in the recording do point to Peshu, so the reporter's own variety leaves some trace in the text, but it loses on every aggregate measure. What the models report is the variety on the recording.

The two recordings separate in the direction the speakers' backgrounds would suggest. The native Thang broadcast is the more confident on every measure available, at 0.974 against 0.831 on the pooled classifier probability and a perplexity margin of 4.07 against 2.13. In the non-native broadcast the mass Thang gives up is split between Nokhu, the variety the reporter married into, which rises from 0.228 to 0.270, and Peshu, which rises from 0.042 to 0.106 and stays last. This is the pattern accommodation would produce.

The intervals holding no Khiamniungan text, 12 in the first recording and 11 in the second, were passed over un-scored, and in both the text stops early: the last transcribed interval ends at 97.9 seconds in the first recording and at 57.0 in the second, against broadcasts running 615 and 591 seconds. Both verdicts are therefore drawn from the opening minute and a half and not from a sample spread across the broadcast, and the great majority of each broadcast carries no transcript at all. Roughly half a minute of Khiamniungan-only automatic transcript was enough for both models to agree on the right dialect twice, and that is the scale at which the pipeline was tested here.

The per-interval scores show one error pattern worth naming. Each recording has intervals the classifier gives to Nokhu, and in both cases the strongest such interval is the opening announcement: *ʔa hʊ̈ me nɔŋ ʔɔ lin ɲae tʊ̈ kɔ hi ma* at 18.2s in the first recording, taken as Nokhu at p = 0.853, and *nɔŋ ʔa ka ʃi wa ni kɔ hi ma* at 20.8s in the second, taken as Nokhu at p = 0.797. The two share nɔŋ and the closing kɔ hi ma, and the classifier's own accounting puts *hi* and *kɔ ʔa* among the features arguing hardest against Thang in the first recording, hi again in the second, and *nɔŋ* in both lists lower down. Formulaic broadcast openings recur across reporters and dialects, so the per-interval errors in these two recordings are not independent of each other, and per-interval accuracy on radio material will be held down by material that is not dialectal in the first place.

5 Conclusion and Outlook

A larger corpus of expertly transcribed speech covering a wider range of genres would allow for more accuracy in identification. Having recordings from additional sub-dialectal varieties of the three main branches would likewise improve the results as it would allow for village-specific identification in cases where there are village-specific markers. Khiamniungan has such markers, with each village having small differences in phonology and lexicon, and each clan within a village having such differences that are discernible to native speakers. Whether these are purely phonological is yet to be determined experimentally.

How much transcribed speech a new variety needs is not established here. Thaam et al. (2026) give the equivalent curve for the transcription stage, where phone error rate flattens within three minutes of connected speech added to the seed wordlist and the twelve minutes after that buy nothing further. The dialect stage has no such curve. Plotting identification accuracy against minutes transcribed per variety would say what it costs to extend the method to a language at the beginning of documentation, which is the question a documenter deciding whether to attempt this will ask first.

The 51 markers of §4.2 are statistical findings and we have not checked them against the isoglosses already described for Khiamniungan. Establishing which of them recover known dialect differences and which are new candidates is descriptive work of a different kind from the question this paper asks, and we leave it to a study with that focus. The results of this study show that even with a small amount of text-only data and early-stage ASR models, dialect identification can be accomplished efficiently and with minimal computing power.

On the six-speaker corpus both models label 37 of the 38 held-out files correctly, at 0.974 accuracy and 0.975 macro F1, and the classifier reaches 0.791 macro F1 scoring single intervals of a few seconds each. Two broadcasts outside every training set, one of them by a reporter whose home variety is not the one on air, were labelled correctly by both models from about half a minute of automatic transcript each.

The computing power claim can be given a figure on each side. For the transcription stage, the cost was paid once by Thaam et al. (2026), who report a training run of under an hour on a single consumer GPU; we run that model in inference only and add nothing to it. The dialect stage costs far less than that. It holds no neural network and touches no GPU: the classifier is one L2-penalized logistic regression over a sparse feature matrix of 3,412 rows, the generative models are order-5 character n-gram models with Witten-Bell interpolation, and the fitted bundle that carries both, together with the vectorizers, the markers and the report, is 0.86 MB on disk. A project that already has a working ASR model can add dialect identification to it on the machine it edits transcripts on.

References

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird & Alexis Michaud (2018): Evaluation Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki: European Language Resources Association.

Agresti, Alan (2001): Exact inference for categorical data: recent advances and continuing controversies. Statistics in Medicine 20. 2709-2722. DOI: https://doi.org/10.1002/sim.738

Ali, Ahmed, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell & Steve Renals (2016): Automatic Dialect Detection in Arabic Broadcast Speech. In: Interspeech 2016. 2934-2938. DOI: https://doi.org/10.21437/Interspeech.2016-1297

Anderson, Cormac, Tiago Tresoldi, Thiago C. Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel & Johann-Mattis List (2018): A Cross-Linguistic Database of Phonetic Transcription Systems. Yearbook of the Poznań Linguistic Meeting 4.1. 21-53. DOI: https://doi.org/10.2478/yplm-2018-0002

Benjamini, Yoav & Yosef Hochberg (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 57.1. 289-300. DOI: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Clements, George N. (1990): The role of the sonority cycle in core syllabification. In: John Kingston & Mary E. Beckman (eds.), Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. Cambridge: Cambridge University Press. 283-333.

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed & Michael Auli (2021): Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In: Interspeech 2021. 2426-2430.

Dam, Kellen Parker van (2026, to appear): Patkaian (Northern Naga). In: Kristine Hildebrandt, Yankee Modi, David Peterson & Hiroyuki Suzuki (eds.), The Oxford Guide to the Tibeto-Burman Languages. Oxford: Oxford University Press. DOI: https://doi.org/10.5281/zenodo.22253033

Dam, Kellen Parker van & Keen Thaam (2026a, to appear): The hidden morphology of Khiamniungan: Uncovering once-productive gender marking & augmentatives. In: Dhiren Singha, Mary Kim Haokip & Monali Longmailai (eds.), Under-Documented Languages of Northeast India, vol. 2. Delhi: Mittal Publishers.

Felsenstein, Joseph (1985): Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39.4. 783-791. DOI: https://doi.org/10.1111/j.1558-5646.1985.tb00420.x

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke & Travis E. Oliphant (2020): Array programming with NumPy. Nature 585. 357-362. DOI: https://doi.org/10.1038/s41586-020-2649-2

Hunter, John D. (2007): Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9.3. 90-95. DOI: https://doi.org/10.1109/MSCE.2007.55

Jauhiainen, Tommi, Marco Lui, Marcos Zampieri, Timothy Baldwin & Krister Lindén (2019): Automatic Language Identification in Texts: A Survey. Journal of Artificial Intelligence Research 65. 675-782. DOI: https://doi.org/10.1613/jair.1.11675

Malmasi, Shervin & Marcos Zampieri (2017): German Dialect Identification in Interview Transcriptions. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). Valencia: Association for Computational Linguistics. 164-169. DOI: https://doi.org/10.18653/v1/W17-1220

Michaud, Alexis, Oliver Adams, Trevor Anthony Cohn, Graham Neubig & Séverine Guillaume (2018): Integrating Automatic Transcription into the Language Documentation Workflow: Experiments with Na Data and the Persephone Toolkit. Language Documentation & Conservation 12. 393-429.

Monroe, Burt L., Michael P. Colaresi & Kevin M. Quinn (2008): Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Political Analysis 16.4. 372-403. DOI: https://doi.org/10.1093/pan/mpn018

Nerbonne, John & Wilbert Heeringa (1997): Measuring Dialect Distance Phonetically. In: Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97). 11-18.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay (2011): Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12. 2825-2830.

Salton, Gerard & Christopher Buckley (1988): Term-weighting approaches in automatic text retrieval. Information Processing & Management 24.5. 513-523. DOI: https://doi.org/10.1016/0306-4573(88)90021-0

Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey & Sanjeev Khudanpur (2018): X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: https://doi.org/10.1109/ICASSP.2018.8461375

Sokal, Robert R. & Charles D. Michener (1958): A Statistical Method for Evaluating Systematic Relationships. University of Kansas Science Bulletin 38.22. 1409-1438.

Thaam, Keen (2025): Reduplication in Khiamniungan. Himalayan Linguistics 24.2. 79-91. DOI: https://doi.org/10.5070/H9.47049

Thaam, Keen & Mimi Kevichüsa-Ezung (2024): Phonology of Khiamniungan. Vāk Manthan 9.2. 39-47.

Thaam, Keen, Kellen Parker van Dam & David Snee (2026): Bootstrapping Speech Recognition for Khiamniungan, an Under-Documented Tibeto-Burman Language. Preprint, not peer reviewed: https://openreview.net/forum?id=3TNj8LPT0D

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa & Paul van Mulbregt (2020): SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17. 261-272. DOI: 10.1038/s41592-019-0686-2

Wieling, Martijn & John Nerbonne (2015): Advances in Dialectometry. Annual Review of Linguistics 1. 243-264. DOI: https://doi.org/10.1146/annurev-linguist-030514-124930

Witten, Ian H. & Timothy C. Bell (1991): The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory 37.4. 1085-1094. DOI: https://doi.org/10.1109/18.87000

Zampieri, Marcos, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer & Noëmi Aepli (2017): Findings of the VarDial Evaluation Campaign 2017. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). Valencia: Association for Computational Linguistics. 1-15. DOI: https://doi.org/10.18653/v1/W17-1201

Zissman, Marc A. (1996): Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. IEEE Transactions on Speech and Audio Processing 4.1. 31-44. DOI: https://doi.org/10.1109/TSA.1996.481450

Supplementary Material

The supplementary material accompanying this study has been curated on Codeberg (https://codeberg.org/besra/supplementary_code). The broadcasts are public transmissions of AIR Akashvani Radio Kohima. Rights in the audio remain with the broadcaster, so we do not redistribute the recordings. The reporters broadcast under their own names, so those names are public, but we do not print them. The dialect labels still attach a village-level identification to individuals who can be identified from the broadcasts; those labels come from information the rater already held about the reporters, with no approach made to the reporters themselves. The wider pipeline takes audio to a village-level or clan-level attribution with no human in the loop, and Khiamniungan is spoken across the India-Myanmar border in an area where an automatic claim about a speaker's origin is not a neutral act. Our intended use is descriptive: sorting an archive of recordings by variety so that documentary material can be found and compared. The released model distinguishes three dialects and offers no finer resolution, and we release no material that names a speaker. We would ask that the same restraint apply to derived models, which is a request and not something a licence can enforce.