Automated Quality Control for Language Documentation: Detecting Phonotactic Inconsistencies in a Kokborok Wordlist

Kellen Parker van Dam[1] **and Abishek Stephen**[2]
[1]Chair for Multilingual Computational Linguistics, University of Passau, Germany
[2]Institute of Formal and Applied Linguistics, Charles University, Czech Republic

Abstract

Lexical data collection in language documentation often contains transcription errors and borrowings that can mislead linguistic analysis. We present unsupervised methods to identify phonotactic inconsistencies in wordlists, applying them to a multilingual dataset of Kokborok varieties with Bangla. Using phoneme-level and syllable-level n-gram language models, our approach identifies potential transcription errors and borrowings. We evaluate our methods using hand annotated gold standard and rank the phonotactic outliers using precision and recall at K metric. The ranking approach provides field linguists with a method to flag entries requiring verification, supporting data quality improvement in low-resourced language documentation.

1 Introduction

In linguistic fieldwork, description frequently begins with the collection of lexical data (Chelliah, 2014). This is often done by means of concept lists such as those of Swadesh (Swadesh, 1955). In the early stages of research, initial elicitation sessions commonly produce "messy" data. When lexical items are collected as a preliminary step, the fieldworker may not be fully acquainted with the phonological system of the target language. Which sounds are phonemic as opposed to allophonic variation may be uncertain. As a result, words are often transcribed narrowly which may be inconsistent from entry to entry regarding the underlying phonemes, as well as potentially failing to capture the underlying phonemic contrasts. A lack of systematicity in the transcription can in turn create issues later on in data analysis (Himmelmann, 1998).

For proper documentation it is important to incorporate the speech of multiple participants. However, when data is drawn from multiple speakers in this manner, differences in dialect or accent may be reflected in the forms recorded. Variation between careful and casual speech styles can introduce further irregularities (Chelliah, 2014).

Lexical borrowing constitutes another potential complication. Borrowed terms may have varying degrees of adherence to the underlying phonemic system. Terms may also have been borrowed twice, with an intermediate borrowing of another closely related language having a different set of phonotactic constraints, thus obscuring their borrowed nature. Thus, it is important that borrowings can be readily identified when attempting to understand the phonology of a language. Borrowed forms may enter the dataset without the researcher's awareness, particularly when the donor language is unfamiliar to the fieldworker.

The context of elicitation is also important. Differences in approaches can exert a significant influence on the quality and consistency of the data. The degree of formality in the interaction, the presence of other speakers, and the level of fatigue or attention on the part of the consultant can all affect the data. The fieldworker's own background and expectations also shape the data in subtle but consequential ways (Kelly and Lahaussois, 2021).

For these reasons, having a method for detection of phonological outliers is of great value to the documentary linguist. By identifying potential borrowings or inconsistencies introduced by factors, the end result of any descriptive study is immediately aided in the very first steps of lexical data collection. Having automated flags for "this entry looks phonotactically weird" could save field linguists considerable time, especially when working with under-resourced languages where you can't rely on external data verification. We do this detection using n-gram language modeling based on phoneme and syllable-level analysis.

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 1–7
March 26, 2026.

2 Related Work

The current research deals with identifying the phonotactic inconsistencies in a linguistic wordlist which in a different light can be seen to have concordances with spelling checkers or borrowing detection methods. Our work however, is not aimed towards either of them albeit the overt similarities. Worth mentioning are some attempts of borrowing detection using wordlists. Miller et al. (2021) where automatic methods for detecting lexical borrowings from monolingual wordlists, comparing different neural network based architectures. List (2019) presents approaches for detecting language contact and borrowing, focusing on phylogenetic networks, sequence comparison methods for detecting borrowings in multilingual wordlists, and trait-based approaches that distinguish borrowed from inherited features using borrowability arguments.

3 Source Data

We rely on data for Kokborok (Glottocode: tipp1238, Hammarström et al., 2025), an underdescribed Tibeto-Burman language group under the Barish language branch (Delancey, forthcoming).

The consonant inventory is relatively moderate in size, with a notable series of aspirated stops that likely developed through Indo-Aryan influence, as aspiration contrasts are less common in many Tibeto-Burman languages. The language maintains voicing distinctions across bilabial, dental, velar, and palatal places of articulation. Word-finally, however, obstruents typically devoice, a pattern not found in neighboring Bangla. Notably, voiced affricates like /dʒ/ are not native to Kokborok but appear in Bangla loanwords, representing sounds borrowed along with vocabulary. Kokborok strongly prefers open syllables and avoids consonant clusters, reflecting its Tibeto-Burman phonotactic constraints. This contrasts sharply with Bangla, which permits complex consonant clusters both word-initially and word-finally. Where Bangla allows syllables like /bdʒro/ or final clusters like /-sto/, Kokborok maintains simpler CV(C) structures with very limited coda positions.

These phonotactic differences create a clear phonological boundary between Kokborok and Bangla despite their geographic proximity. The result is two typologically distinct systems coexisting in close contact, with Kokborok maintaining its characteristic Tibeto-Burman simplicity in syllable structure even while absorbing lexical material from its Indo-Aryan neighbor.

Our data comes from Kim et al., 2025, a sociolinguistic survey of 306 concepts in 20 Kokborok varieties plus 3 varieties of Garo (garo1247), and standard Bangla as the main contact language. This concept list is a good representation of the language as it covers the majority of basic morphemes which go into lexical construction. As is typical in Tibeto-Burman languages of the region, words are primarily compounds of simpler common morphemes. This set of 306 concepts is considerably larger than the amount that would normally go into computational phylogenetic work, such as the 180 concepts of Sagart et al. (2019) or the 100 of Galucio et al. (2015), and coverts the full range of phonological variation occurring in native forms.

Kokborok data were converted to the Cross-Linguistic Data Format (CLDF; Forkel et al. (2018)) by the authors. Morphological features external to the citation form were removed, as were erroneous repeated diacritics which did not contribute to the transcription. The data were otherwise not modified, leaving ambiguous transcriptions[1] as is. The CLDF dataset is available in a GitHub repository[2]

4 Experiments

The identification of the phonological anomalies or outliers in our dataset proceeds via implementing simple n-gram language models at the phoneme and syllable levels. First, we run the experiments on the phoneme level which aims to capture rare phoneme sequences and then we contrast it with positional phonotactics using the syllable-level analysis that captures more linguistically motivated violations. The source codes are publicly available[3].

The training data has 3055 words after removing duplicates[4]. The gold data contains 555 words marked as borrowings. Our annotations focus exclusively on borrowings, as these were straightforward to identify given the clear phonological and lexical distinctions between Kokborok and Bangla, the primary source of loanwords in the dataset. We treat words in Kokborok that are almost identical to the

[1] For example in the AbengGajni doculect, the verb 'to eat' is erroneously transcribed as tshaʔ with no clear indication if this should be tʃaʔ or tsʰaʔ.

[2] https://github.com/phonemica/kimkokborok. The dataset can also be accessed here-https://doi.org/10.5281/zenodo.17973867.

[3] https://github.com/abishekjs/kokborok-anotect

[4] Since the linguistic varieties or doculects are closely related, the same word forms are used to encode a given semantic concept.

Bangla counterpart phonologically (as explained in § 3) for a given semantic concept as borrowed. For example, the word for 'rainbow' in Bangla and doculect MukchakBarbakpur is ɾɔŋdʰɔnu and hence marked as a borrowing. Transcription errors were not systematically annotated due to the difficulty of distinguishing genuine errors from dialectal variation or unknown phonological processes, making borrowings more reliable for evaluating our methods.

4.1 Phoneme-level N-gram Language Modeling

To identify phonotactic anomalies such as transcription errors and borrowings, we train phoneme bigram and trigram language models on the Kokborok data. Words are padded with boundary markers (e.g., ^noukʰa$), and diacritics are treated as separate characters. We apply Laplace smoothing to handle unseen n-grams.

Our mathematical assumption is that the words with transcription errors or borrowings would have some phoneme sequences which are rare in the language, and using the negative log likelihood (NLL) such words would be flagged when ranking by resulting NLL scores. This can help linguists make quick quality checks, as the stronger outliers would be captured in the top K words. We compute NLL for words using different aggregation methods to capture character-level variations.

Arithmetic Mean captures the expected information content per n-gram. It normalizes for word length enabling fair comparisons between long and short words.

$$\text{Mean NLL} = -\frac{1}{N}\sum_{i=1}^{N}\log_2 p(x_i) \qquad (1)$$

where N is the number of n-grams in the word and x_i represents the i-th n-gram.

Harmonic Mean emphasizes the most typical n-grams within a word, being heavily weighted toward smaller NLL values. This metric is particularly useful for identifying words that contain a core of native phonotactic patterns even when some unusual or rare n-grams are present.

Min identifies the most typical n-gram in a word, revealing whether the word shares any common phonotactic patterns with the native vocabulary. Even heavily borrowed words may contain some typical n-grams, and this metric helps assess the degree of phonotactic integration of loanwords into the native system. A very low minimum NLL suggests the word has at least partial structural overlap with native phonotactics.

Max identifies the most atypical n-gram in a word, a rare n-gram can be a reminiscent of the source language in case of the word being borrowed. It could also potentially flag off partially integrated loanwords.

4.2 Syllable-level N-gram Language Modeling

We implement automatic syllabification based on the sonority hierarchy and maximum onset principle, where syllable boundaries are determined by identifying sonority peaks and applying language-universal syllable structure constraints. The syllable boundary symbol (.) serves as a structural marker. Here too, we add boundary markers and use the aggregation methods used in the phoneme level analysis.

4.2.1 Analysis Types

We employ three distinct approaches to analyze phonotactic patterns:

- **Within-syllable analysis**: We calculate negative log likelihood of character n-grams that occur strictly within individual syllables, respecting syllable boundaries and focusing on internal syllabic structure.
- **Cross-boundary analysis**: We extract character n-grams that span across syllable boundaries, capturing phonotactic patterns that violate typical syllable constraints and may indicate borrowing or transcription anomalies.
- **Boundary-as-phoneme analysis**: We treat syllable boundaries as legitimate phonemes in the sequence, allowing n-grams to include the syllable boundary symbol and capturing positional sensitivity at syllable edges.

4.3 Results

For the field linguists, it would be highly efficient to discover anomalies based on the ranking of the words following their NLL scores observed for all of the aggregation methods. To facilitate that we use recall and precision at K as our evaluation metric. We use the mean NLL as the baseline for the experiments. The gold data is hand annotated, the words

Table 1: Precision and Recall at K for different NLL aggregation methods and baselines for the phoneme-level n-gram models.

N-gram	Method	P@100	P@500	P@1000	R@100	R@500	R@1000
Bigram	Arithmetic Mean	0.43	0.32	0.26	0.08	0.29	0.47
	Harmonic Mean	0.43	0.31	0.26	0.08	0.28	0.47
	Min NLL	0.36	0.32	0.23	0.06	0.29	0.42
	Max NLL	0.32	0.26	0.23	0.06	0.24	0.42
	Uniform Random	0.15	0.19	0.17	0.03	0.17	0.31
	Stratified Random	0.17	0.20	0.17	0.03	0.18	0.30
Trigram	Arithmetic Mean	0.45	0.33	0.28	0.08	0.30	0.51
	Harmonic Mean	0.46	0.32	0.29	0.08	0.28	0.52
	Min NLL	0.40	0.32	0.27	0.07	0.29	0.49
	Max NLL	0.20	0.29	0.25	0.04	0.26	0.46
	Uniform Random	0.20	0.17	0.19	0.04	0.15	0.34
	Stratified Random	0.32	0.21	0.19	0.06	0.19	0.34

borrowed from Bangla are labeled as borrowings. The current dataset do not have any transcription errors, but the assumption and also the strong caveat of our method also ensures the flagging of such errors.

We establish two random sampling baselines to evaluate whether our n-gram phonotactic models perform better than chance. The uniform random baseline samples K words randomly from the wordlist without any prior assumptions, supporting the hypothesis where all words are equally likely to be anomalies. The stratified random baseline samples words proportionally by length, controlling for the possibility that transcription errors or borrowings may be biased toward longer or shorter words.

4.4 Phoneme-level Results

Table 1 demonstrates that our phoneme-level n-gram phonotactic models substantially outperform random baselines in identifying phonotactic anomalies. Trigram models achieve the strongest performance, with precision at 100 reaching 0.46 and recall at 1000 reaching 0.52 using harmonic mean aggregation. The superiority of trigrams over bigrams suggests that richer phonotactic context is crucial for capturing constraints violations, while the consistent performance of arithmetic and harmonic mean aggregation indicates that anomalies are characterized by sustained phonotactic unusualness across the entire word rather than isolated rare n-grams. At K=500, our best model achieves 33% precision, meaning that approximately one in three flagged items is a genuine anomaly. However, the plateau at 52% recall suggests that roughly half of the gold anomalies are phonotactically well-formed, indicating they may represent semantic borrowings or transcription errors that do not violate native phonological constraints.

The best performing model based on trigram-harmonic mean identifies words like ʤʰaɽu, oʃud, ʈɪkʈɪki, ɔnɛk, mɛgʰgɔɹʤon and so on in the top 100 words being flagged as anomalous. In the top 500 words like poɾibar, murgi, gɔrom get flagged. This ranking pattern reflects the model's sensitivity to different degrees of phonotactic deviations, top 100 of the flagged words typically contain phonemes or phoneme combinations that are extremely rare or absent in native Kokborok vocabulary (such as retroflex consonants and aspirated affricates), while words ranked in the top 500 exhibit more subtle violations involving less frequent but attested phoneme sequences, suggesting partial phonological adaptation common for borrowings.

4.5 Syllable-level Results

Our results (Table 3, see Appendix) reveal distinct performance patterns across three phonotactic modeling approaches. Within-word analysis achieves the strongest overall performance with precision at 100 of 0.47 for bigrams and recall at 1000 of 0.49 for trigrams, effectively capturing internal phonological structure violations. Boundary-as-phoneme analysis shows competitive results, particularly at lower K values where trigram models reach precision at 100 of 0.49, indicating that syllable boundary constraint violations are highly predictive of anomalies. In contrast, cross-boundary analysis substantially underperforms, with precision rarely exceeding 0.30 and recall at 1000 capped at 0.42, suggesting that phonotactic violations are better characterized by position-specific patterns

Table 2: Precision and Recall at K for bigram models with arithmetic mean aggregation across different syllable analysis types.

Analysis	P@100	P@500	P@1000	R@100	R@500	R@1000
Within	0.47	0.32	0.26	0.08	0.29	0.47
Cross	0.21	0.20	0.18	0.04	0.18	0.33
Boundary	0.50	0.29	0.24	0.09	0.26	0.43
Uniform Random	0.15	0.19	0.17	0.03	0.17	0.31
Stratified Random	0.17	0.20	0.17	0.03	0.18	0.30

within syllable constituents rather than by boundary-crossing transitions alone.

Words like mɛgʰ and morɪʧ are caught early on at top 100 using the boundary-as-phoneme bigram arithmetic mean setup (Table 2). These words were flagged off in the top 500 of the phoneme trigram harmonic mean setup. This demonstrates the complementary strengths of syllable-aware modeling. The boundary marker (.) itself being present in some sequence is highly influential on results in that it closely reflects syllable position i.e. the phonemes preceding (.) indicate syllable onset or nucleus, following (.) indicates coda position and so on.

5 Conclusion

Our study addresses transcription errors and unidentified borrowings that can skew typological analysis in wordlist-based language documentation. Designed for the initial stages of data collection, these methods provide field linguists with systematic tools to identify entries requiring closer inspection. Our results reveal that phoneme-level n-gram models capture most anomalies also flagged by syllable-level models, suggesting that while incorporating explicit phonotactic knowledge through syllabification provides some benefit, raw phoneme sequence modeling alone achieves comparable performance. This indicates that computationally simpler approaches may be sufficient for practical anomaly detection in fieldwork settings, though the syllable-level analysis offers additional interpretability by identifying specific constraint violations.

Limitations

Due to the limited size of documentary wordlists, data-intensive approaches such as neural language models cannot be applied, though such methods might yield superior performance in high-resource settings. Borrowing detection is inherently challenging making gold standard creation labor-intensive. However, the statistical nature of n-gram models ensures they capture phonotactic inconsistencies providing field linguists with a tool for identifying entries that deviate from expected patterns and require closer inspection.

In terms of the method's usefulness in borrowing detection, this relies heavily on having a phonotactically more restrictive language borrowing from one with a greater possibility of sounds, or simply sounds which are not found in the borrowing language. Detecting Kra-Dai borrowings into a phonologically similar Tibeto-Burman language would not be possible in this case. However, the method is still useful in detecting transcription errors, novel sound changes, dialectal variation, or other cases which still warrant further investigation by the linguist even if not the result of borrowing.

Finally, as the fieldworker's expectations also shape how they transcribe data (Kelly and Lahaussois, 2021), application of this approach too quickly or without further investigation into the language could result in further over-application of mistakes. One should not blindly assume anomalies are errors. Rather, they are points to be investigated further and confirmed.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was supported by the Charles University student project GA UK No. 101924 and partially supported by SVV project number 260 698.

References

Shobhana Chelliah. 2014. Fieldwork for language description. *Research methods in linguistics*, pages 51–73.

Scott Delancey. forthcoming. The barish languages. In K. Hildebrandt, Y. Modi, D. Peterson, and H. Suzuky, editors, *The Oxford Guide to the Tibeto-Burman Languages*. Oxford University Press, Oxford.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and reuse in comparative linguistics. *Scientific Data*, 5:180205.

Ana Vilacy Galucio, Sérgio Meira, Joshua Birchall, Denny Moore, Nilson Gabas Júnior, Sebastian Drude, Luciana Storto, Gessiane Picanço, and Carmen Reis Rodrigues. 2015. Genealogical relations and lexical distances within the tupian linguistic family. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, 10:229–274.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. *Glottolog 5.2*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available online at http://glottolog.org, Accessed on 2025-08-08.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Barbara Kelly and Aimée Lahaussois. 2021. Chains of influence in Himalayan grammars: Models and interrelations shaping descriptions of Tibeto-Burman languages of Nepal. *Linguistics*, 59(1):207–245.

Amy Kim, Palash Roy, Mridul Sangma, and Seung Kim. 2025. Cldf dataset derived from kim et al's "the tripura of bangladesh: A sociolinguistic survey" from 2011. Version v1.0.0, published December 18, 2025.

Johann-Mattis List. 2019. Automated methods for the investigation of language contact, with a focus on lexical borrowing. *Language and Linguistics Compass*, 13(10):e12355.

John Miller, Emanuel Pariasca, and Cesar Beltran Castañon. 2021. Neural borrowing detection with monolingual lexical models. In *Proceedings of the student research workshop associated with RANLP 2021*, pages 109–117.

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin J Ryder, Valentin Thouzeau, Simon J Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of sino-tibetan. *Proceedings of the National Academy of Sciences*, 116(21):10317–10322.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

A Appendix

Table 3: Precision and Recall at K for different analysis types and n-gram sizes for the syllable-level n-gram models.

Analysis	N-gram	Method	P@100	P@500	P@1000	R@100	R@500	R@1000
Within	Bigram	Arithmetic Mean	0.47	0.32	0.26	0.08	0.29	0.47
		Harmonic Mean	0.43	0.32	0.26	0.08	0.29	0.48
		Min NLL	0.35	0.32	0.23	0.06	0.29	0.42
		Max NLL	0.35	0.28	0.23	0.06	0.26	0.41
	Trigram	Arithmetic Mean	0.45	0.31	0.27	0.08	0.28	0.49
		Harmonic Mean	0.45	0.31	0.27	0.08	0.28	0.48
		Min NLL	0.40	0.32	0.27	0.07	0.29	0.49
		Max NLL	0.37	0.29	0.23	0.07	0.26	0.42
Cross	Bigram	Arithmetic Mean	0.21	0.20	0.18	0.04	0.18	0.33
		Harmonic Mean	0.21	0.20	0.19	0.04	0.18	0.33
		Min NLL	0.22	0.19	0.17	0.04	0.17	0.30
		Max NLL	0.34	0.21	0.19	0.06	0.19	0.34
	Trigram	Arithmetic Mean	0.30	0.26	0.21	0.05	0.23	0.37
		Harmonic Mean	0.30	0.26	0.21	0.05	0.23	0.37
		Min NLL	0.30	0.26	0.21	0.05	0.24	0.38
		Max NLL	0.26	0.27	0.23	0.05	0.24	0.42
Boundary	Bigram	Arithmetic Mean	0.50	0.29	0.24	0.09	0.26	0.43
		Harmonic Mean	0.44	0.30	0.25	0.08	0.27	0.45
		Min NLL	0.41	0.25	0.18	0.07	0.22	0.33
		Max NLL	0.36	0.28	0.23	0.06	0.26	0.42
	Trigram	Arithmetic Mean	0.48	0.29	0.26	0.09	0.26	0.47
		Harmonic Mean	0.49	0.29	0.26	0.09	0.26	0.47
		Min NLL	0.44	0.29	0.25	0.08	0.26	0.46
		Max NLL	0.37	0.30	0.24	0.07	0.27	0.44
Baseline	Uniform Random		0.15	0.19	0.17	0.03	0.17	0.31
	Stratified Random		0.17	0.20	0.17	0.03	0.18	0.30