Statistical and Neural Methods for Hawaiian Orthography Modernization

Jaden Kapali **Keaton Williamson** **Winston Wu**
Department of Computer Science
University of Hawai‘i at Hilo
{jkapali,georgekw,wswu}@hawaii.edu

Abstract

Hawaiian orthography employs two distinct spelling systems, both of which are used by communities of speakers today. These two spelling systems are distinguished by the presence of the ‘okina letter and kahakō diacritic, which represent glottal stops and long vowels, respectively. We develop several models ranging in complexity to convert between these two orthographies. Our results demonstrate that simple statistical n-gram models surprisingly outperform neural seq2seq models and LLMs, highlighting the potential for traditional machine learning approaches in a low-resource setting.

1 Introduction

The Hawaiian language (‘Ōlelo Hawai‘i) is the indigenous language of Hawai‘i, an island chain in the central Pacific consisting of eight main islands. The Hawaiian language did not have a written form until the 1820s when Protestant missionaries from New England introduced a writing system using the Latin alphabet. This early orthography represented an ambiguous one-to-many relationship between spelling and sound. The orthography did not account for differences in vowel length and the presence of the glottal stop (ʔ), common in Austronesian languages.

Between the 1930s and 1950s, Mary Kawena Pukui, a legendary Hawaiian scholar and cultural preservationist, kept detailed lists of Hawaiian words during her work at the Bishop Museum. In 1949, she was asked by Sir Peter Buck, then director of the Bishop Museum, to author a new Hawaiian-English dictionary (Lam, 1957). In 1952, Samuel Elbert, a linguistics professor and long-time student and friend of Pukui, began phonetically systematizing Pukui's vast word lists (Handy and Pukui, 1972). This group effort culminated in the Pukui-Elbert Hawaiian Dictionary (Pukui and Elbert, 1957), which introduced a new spelling system with additional characters: the ‘okina, which represents a glottal stop (ʔ), and kahakō, which mark long vowels (ā ē ī ō ū). With these modifications, the spelling system provided a nearly one-to-one mapping between sound and orthography, helping to disambiguate words, as shown in Table 1. These modifications to the orthography aimed to preserve the pronunciation of Hawaiian words for future generations, and this spelling system quickly became the preference of a new generation of language learners during the Hawaiian Renaissance of the 1960s and 1970s.

Although there were likely many dialects across the island chain at one time, today there are two main groups of speakers that have perpetuated and grown the use of ‘Ōlelo Hawai‘i over the years, despite the challenges of more than 100 years of language decline in Hawai‘i. **Olelo Niihau** is a dialect unique to the island of Niihau that has persisted due to the isolation of Niihau as a privately owned island since 1864. The Niihau community is the last intact community of native speakers of traditional Hawaiian in the world (Wong et al., 2025). **‘Ōlelo Hawai‘i** describes all other dialects of Hawaiian. Today, there is an increasing number of speakers of ‘Ōlelo Hawai‘i thanks to the establishment of Hawaiian-medium education schools in the 1980s that have grown in popularity since then.

Both varieties of Hawaiian have grown and evolved over time, with differences in spelling, pronunciation, and word choice. Their speakers generally prefer two different spellings. Olelo Niihau speakers tend to use the older spelling, which omits the ‘okina and kahakō. Glottal stops and long vowels are instead implied from context, except in pronouns, where glottal stops are always written using an ‘okina or apostrophe to disambiguate homographs (Wong, 2020). In contrast, ‘Ōlelo Hawai‘i speakers tend to use the newer spelling, where the ‘okina and kahakō are always written where they occur, and the ‘okina is considered a letter of the

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35138–35144
November 4-9, 2025

alphabet. The newer spelling is the predominant system taught in schools today. For the purposes of this paper, the earlier spelling system, used by previous generations and by the Niihau community today, will be called **Traditional Spelling** (TS), while the newer spelling system, preferred by most speakers today, will be called **Modern Spelling** (MS). These are not perfect terms — both speech communities are modern, and both uphold tradition in their use of the Hawaiian language.

There are strong motivations for developing systems to convert between these two orthographies. Although MS is preferred by many fluent and L1 Hawaiian speakers today, many speakers of Olelo Niihau, who come from a tradition of reading The Hawaiian Bible (*Ka Baibala Hemolele*) without the ʻokina and kahakō, say that the pronunciation of words in TS is clear, and that MS can even hinder their ability to read (Wong and Faria, 2021). This could be attributed to increased visual complexity or less familiarity with MS, as seen among Arabic readers (Maroun, 2017). In the context of language revitalization in Hawaiʻi, many speakers and learners of Hawaiian rely on the ʻokina and kahakō found in MS for correct pronunciation and meaning. The same trend is seen in Arabic, Samoan, and Cook Islands Māori, where a more phonemic writing system is helpful in language learning contexts (Maroun, 2017; Tualaulelei et al., 2015; Ministry for Pacific Peoples, 2024a,b).

Thus, developing computational methods that can accurately convert between TS and MS would greatly benefit the Hawaiian language community, improving the intelligibility of texts for a wider range of audiences. Our paper presents experiments on this orthography conversion task with a variety of models, showing that extremely simple statistical models perform better than complex neural models commonly used for other NLP tasks. We release code for our experiments at `https://github.com/wswu/hawspell`.

2 Related Work

One of the first to tackle this task was Yarowsky (1994), who used n-gram based methods to restore accents in Spanish and French. Diacritic restoration has been developed for various languages such as Czech (Richter et al., 2012), Turkish (Adali and Eryiğit, 2014), Hungarian (Novák and Siklósi, 2015), and Arabic (Schlippe et al., 2008). Similar to one aspect of our work are neural sequence-to-sequence models for diacritic restoration for languages such as Arabic (Belinkov and Glass, 2015) and Vietnamese (Náplava et al., 2018; Dang and Nguyen, 2020). Similar models have also been used for related tasks such as stress prediction and syllabification (Wu and Yarowsky, 2021). While most existing works employ character-level models, we also experiment with subword tokenization. Furthermore, existing works largely target high-resource languages with hundreds of thousands or even millions of training examples, in contrast to Hawaiian's low-resource nature.

Traditional	Modern	English
kai	kai	sea
kai	kaʻi	march
kai	kāī	a kind of taro
au	au	current
au	āu	your
au	ʻau	swim
au	aʻu	marlin

Table 1: An excerpt of a table from Wilson (1981) showing Hawaiian words in Traditional and Modern spelling.

For Hawaiian, Walker et al. (2025) created an online dictionary platform with support for entries in TS and MS. Shillingford and Parker Jones (2018) use RNNs and finite-state transducers for the task of orthography conversion to MS. We build upon their work, comparing both similar and more complex models, and investigate their hypotheses that neural methods would not perform well on this task due to the small amount of Hawaiian data.

3 Data

We use audio transcripts from Kaniʻāina (Kimura, 2023), a collection of radio and TV programs (*Ka Leo Hawaiʻi* and *Kū i ka Mānaleo*) recorded in the 1970–1980s. The guests in these programs were some of the few remaining native speakers alive at the time. They discussed various topics from daily life, to culture, to language preservation. Thus, the text represents real-world usage of Hawaiian.

We preprocess the transcripts by removing English sentences and boilerplate text (e.g. some transcripts start with music or a sentence like "Beginning of the Hawaiian tape HV24 reel 2"). Because the transcripts are written using MS, we follow Shillingford and Parker Jones (2018) to cre-

# MS Variants	1	2	3	4	5
Words	12,576	886	140	48	14

Table 2: Histogram of the number of unique TS words with MS variants. Most words only have one variant. One word with five variants, *ia*, is shown in Table 3.

Word	Freq	English
ia	7,446	he/she/it, that
iā	4,338	*accusative marker*
‘ia	3,218	*passive marker*
i‘a	355	fish
‘īā	6	yard

Table 3: The word *ia* has several spelling variants in MS. In many cases, the correct form can be inferred from context.

ate synthetic TS text for our task by removing all ‘okina (‘) and replacing all letters with the kahakō (macron) with the bare letter. This process resulted in a total of 76,980 sentences, which we split into 80% train, 10% validation, and 10% test sets.

As a preliminary analysis, we first examine ambiguity (homographs) in TS in the data. As shown in Table 2, most words in TS have only a single MS variant and are unambiguous in both TS and MS. One of the words with the most variants, *ia*, is shown in Table 3. Other words with a high number of MS variants include *ai, aina, ana, ano*, and *ii*.

4 Methods

To convert from MS to TS, we simply remove all ‘okina (‘) and replace letters with the kahakō (ā, ē, ī, ō, ū) with the equivalent letter without the diacritic. This is a deterministic mapping. To convert from TS to MS, we develop several methods:

N-gram. The n-gram method uses Bayes Rule to find the most likely variant given the words in its left context, $c = w_{i-n+1:i-1}$.

$$\hat{v} = \underset{v \in \text{variants}(w_i)}{\text{argmax}} \; p(v|c) = \frac{\text{count}(c, v)}{\text{count}(c)}$$

For unigrams, we simply use the most common MS variant of the TS word in the training data. For example, in our data, *o* would be consistently converted to *‘o*, and *ia* would be left unchanged using this approach. Yarowsky (1994) has shown the unigram approach to be a strong baseline for French and Spanish accent restoration. For higher-order n-grams, context is likely useful for identifying the correct variant v for the TS word w_i. For example, *ia* is more likely to be converted into *‘ia* if it occurs after a verb. We experiment with n ranging from 1 to 5, with a fallback to a lower order n-gram model if the count is zero. Additionally, we experiment with n-gram counts from HawCorpus (Doherty, 2016), which were computed on a large corpus of Hawaiian text.

Sequence-to-sequence models. We frame the orthography conversion task as a sequence-to-sequence machine translation task, where the source language is the Hawaiian text in TS, and the target language is the same text in MS. With this framing, models will take into account contextual information from the entire sentence. We experiment with two popular encoder-decoder models: a 2-layer LSTM with 500-dim embedding size (10M parameters) and a standard 6-layer Transformer with 8 heads and 512 embedding size (44M parameters). We also vary the vocabulary size when tokenizing the source and target text in order to investigate the effect of characters vs. subword tokenization on this task. These models were implemented using the OpenNMT toolkit (Klein et al., 2017).

Large Language Models. We experiment with few-shot prompts using Llama 3.2-1b and Llama 3.1-8b (Touvron et al., 2023) to add the missing diacritics to the Hawaiian text. In addition, we experiment with finetuning Qwen3-4b-instruct-2507 (Qwen Team, 2025) using Unsloth (Han et al., 2023).

5 Experiments

We split our data into 80-10-10 train-validation-test splits. For the unigram and bigram models, we additionally experiment with training on counts from *HawCorpus* (Doherty, 2016). For the sequence-to-sequence models, we perform tokenization using the WordPiece algorithm from HuggingFace Tokenizers, varying the vocabulary size from 250 to 16k. The neural models were trained with early stopping with a patience of 10 on the validation set. We evaluate models' performance using Word Error Rate (WER) and Character Error Rate (CER).

6 Results and Discussion

Table 4 presents a high-level summary of the results, with sample model outputs shown in Table 6. In

Approach	WER	CER
Original Text	31.06	9.84
Unigram	7.11	2.13
Bigram	4.41	1.42
Trigram	4.02	1.28
4-gram	3.98	1.26
5-gram	3.98	1.26
Unigram (HawCorpus)	23.82	8.03
Bigram (HawCorpus)	21.48	7.49
LSTM	11.96	9.65
Transformer	9.33	8.09
LLM 1B 5-shot	164.21	145.42
LLM 1B 10-shot	241.22	224.75
LLM 8B 5-shot	53.78	20.13
LLM 8B 10-shot	50.66	19.95
LLM Finetune	20.91	20.7

Table 4: Word error rate (%) and character error rate (%) of several approaches to the orthography conversion task (lower is better). Original Text indicates that the TS orthography was evaluated as if it were MS. The best LSTM and Transformer models were character-level models.

this section, we analyze the performance of each approach and present some takeaways.

The statistical n-gram models are simple but surprisingly performed best overall. We find that increasing the order of n-grams improves performance, indicating that context is useful for the task. However, there were no additional gains after $n = 4$. For example, in Table 6, the *o* in *ka inoa o ke mele* (the name of the song) was incorrectly converted as *ʻo* (a particle appearing before proper nouns and pronouns) by the unigram model because *ʻo* is the most common variant of *o*; this mistake was corrected by higher-order models.

The n-gram models using counts from HawCorpus (Doherty, 2016) performed poorly, likely because the corpus from which these counts were computed contains older Hawaiian books, so the TS forms are more frequent than the MS forms. Furthermore, we noticed that the corpus counts often contained non-words like *ʻaʻ*, which likely indicate either OCR errors in the corpus, or tokenization issues before computing the n-gram counts.

In Table 6, all n-gram models incorrectly converted *ana* to *ʻana* (a nominalizing particle). In a bigram context, this makes sense, because *himeni ʻana* means ‘singing‘ (as a noun). However, correctly identifying that *ana* should remain *ana* in

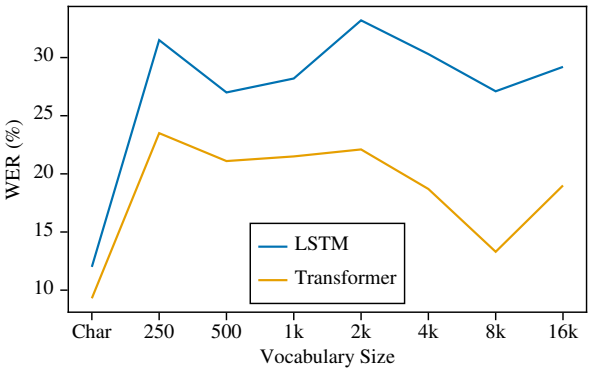

Figure 1: Comparing WER across vocabulary size after tokenization for seq2seq models. Lower is better.

MS requires a larger context: *e <verb> ana* is a pattern that makes a verb future tense, e.g. *e hīmeni ana* (going to sing). The neural seq2seq models and fine-tuned LLM capture this context.

For the neural seq2seq models, overall the Transformer models outperform the LSTM models by a large margin and approach the performance of the n-gram models. We noticed a trend that these models tend to add diacritics to longer words, even if the words do not need them. For example, in Table 6, *akoakoa* is incorrectly converted as *ʼākoʻakoʻa* (coral) rather than *ʻākoakoa* (assemble/gather). Although the Transformer model incorrectly converted the word, *ʼākoʻakoʻa* is indeed an actual Hawaiian word, so we see that the model has learned to predict existing words. On the other hand, it is possible that the smaller LSTM does not have enough parameters to accurately learn the task, as it sometimes hallucinates, e.g. *ʻakoʻakoaʻkoʻa* and *maii* while also rearranging the order of *hōʻike* and *mai*.

We also investigated the effect of subword tokenization on performance, experimenting with character-level tokens and increasing the vocabulary size from 250 to 16k as shown in Figure 1 (the character-level setting corresponds to a source vocabulary size of 136 and a target vocabulary size of 160). Overall, models with data separated into characters performed substantially better than subword-tokenized data. However, there is an improvement at 8k, where most tokens are whole words. This may indicate that tokenization helps correctly convert out-of-vocabulary words, where a portion of the word may have been seen during training. An example is a word that begins with the causative prefix *hoʻo-*, an extremely common subword token, e.g. *hoʻomaʻemaʻe* (to clean). While

subword tokenization is a common preprocessing step in machine translation and other text generation tasks, it may have limited effectiveness for orthography conversion.

Finally, for large language models, we experimented with various prompts for this task. The most successful one was: "Add appropriate ʻokina (ʻ) and kahakō (long vowels) to the following Hawaiian sentence. Only reply with the answer and nothing else. The sentence:" followed by the Hawaiian sentence. When devising prompts, we had to specifically prompt the LLM to not output anything besides the answer. Nevertheless, sometimes the LLM would still output "I can't do that", and a handful of times, it even offered to provide an English translation. To resolve these extraneous outputs, we re-prompted the LLM if its answer contained any English sentences.

Another phenomenon we observed was that, like the neural seq2seq models, sometimes the LLMs would add or remove extra characters. For example, in Table 6, *inoa* was incorrectly converted to *ʻinō*. This seems to be the LLMs rewriting entire words rather than making minimal changes to the spelling. We also found that for short sentences, such as *ʾAe.* or *ʻĒ* (Yes/Yeah), sometimes the model would hallucinate a longer output. Similar to the transformer and LSTM comparison, the larger LLM performed better than the smaller one. The poor performance of LLMs is likely because they have not seen enough Hawaiian during training.

Fine-tuning with sample prompt-response pairs resulted in substantial improvements over the non-fine-tuned models. When finetuning, we did found that it was unnecessary to prompt the model to only reply with the answer and nothing else. We noticed that the fine-tuned model rarely hallucinates new words, and sticks to the task, but tends to generate shorter output. Longer sentences are truncated, resulting in higher WER, but the generated output largely contains correct orthography, as can be seen in Table 6. Future work can experiment with longer sentences or even paragraphs to mitigate this issue.

6.1 Error Analysis

We examined errors across systems and found that *the vast majority involve high-frequency function words* (particles, adpositions, determiners, and discourse markers), where the presence or absence of the ʻokina and kahakō is highly context-dependent. Table 5 lists the ten most frequent errors for the bigram model.

Gold	Predicted	Type	Freq
o	*ʻo*	ʻokina insertion	338
ʻŌ	*ʻO*	kahakō deletion	293
ʻē	*e*	ʻokina/kahakō deletion	192
no	*nō*	kahakō insertion	173
ʻo	*o*	ʻokina deletion	173
ana	*ʻana*	nominalizer over-mark	159
ē	*e*	kahakō deletion	154
kā	*ka*	kahakō deletion	109
E	*ʻĒ*	ʻokina/kahakō insertion	80
na	*nā*	kahakō insertion	71

Table 5: Top 10 errors for the bigram model on the test set. Errors cluster on particles/adpositions and short function words.

Most remaining errors are small, local edits: macron insertions/deletions on vowels in clusters (e.g., ai/ae/ao), and misplaced or missing ʻokina around morpheme boundaries and reduplications. Proper names, loanwords, and homographs whose diacritics depend on lexical knowledge (e.g., pau vs paʻu) are frequent offenders. Normalization mismatches between the true ʻokina and ASCII apostrophe (') also cause extra insertions/deletions. For the best-performing models, the gap between WER ($\tilde{4}$%) and CER ($\tilde{1}$.3%) shows that models typically change only one or two characters per word rather than substituting whole words.

In sum, we have found that simple statistical models outperform more complex neural models for converting between traditional and modern Hawaiian orthography. Models that take into account a larger context window are beneficial for this task. Transformer seq2seq models approach the performance of n-gram based models, though there is still room for improvement. LLMs likely have not seen enough Hawaiian text to be successful at this task, though with additional pretraining or finetuning, they may become viable in the future.

Acknowledgments

This work is partially supported by the National Science Foundation (Award No. 2422413). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. The authors would also like to mahalo the following individuals who provided valuable insights, feedback, and discussion: Kaiapuni Roseguo, Yumiko Ohara, Scott Saft, Pila Wilson, Larry Kimura, Sunny Walker, Bruce Torres Fischer, and Annette Kuuipolani Kanahele Wong.

Limitations

Due to resource constraints (a single NVIDIA A6000 GPU), we could not experiment with larger LLMs. However, we expect that these models would do poorly on this task, because Hawaiian has so little web presence. Anecdotally, we have found through interactions with several leading LLMS that they only understand basic Hawaiian vocabulary and cannot generate more complex phrases or advanced vocabulary. In addition, our current evaluation relies primarily on synthetic proxies for TS. In the future, we plan to assemble and evaluate on a curated set of authentic pre-1950s documents authored originally in TS, with manually-annotated MS.

References

Kübra Adali and Gülşen Eryiğit. 2014. Vowel and diacritic restoration for social media texts. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 53–61, Gothenburg, Sweden. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2015. Arabic diacritization with recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2281–2285, Lisbon, Portugal. Association for Computational Linguistics.

Trung Duc Anh Dang and Thi Thu Trang Nguyen. 2020. TDP – a hybrid diacritic restoration with transformer decoder. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 76–83, Hanoi, Vietnam. Association for Computational Linguistics.

Liam Doherty. 2016. The Hawaiian corpus project: Data from a corpus of written Hawaiian.

Daniel Han, Michael Han, and Unsloth team. 2023. Unsloth.

E. S. Craighill Handy and Mary Kawena Pukui. 1972. *The Polynesian Family System in Kau, Hawaii*.

Larry Kimura. 2023. Kani‘āina, voices of the land.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Jeanette Lam. 1957. "Words are my business," says kamaaina author. *Honolulu Star-Bulletin*, 46(255).

Maryse Maroun. 2017. Diacritics and the resolution of ambiguity in reading Arabic. *University of Essex*, pages 2, 26.

Ministry for Pacific Peoples. 2024a. Gagana Sāmoa orthography guidelines.

Ministry for Pacific Peoples. 2024b. *Te reo Māori Kūki ’Airani Orthography Guidelines.* Wellington, New Zealand.

Jakub Náplava, Milan Straka, Pavel Straňák, and Jan Hajič. 2018. Diacritics restoration using neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Attila Novák and Borbála Siklósi. 2015. Automatic diacritics restoration for Hungarian. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2286–2291, Lisbon, Portugal. Association for Computational Linguistics.

Mary Kawena Pukui and Samuel H Elbert. 1957. *Hawaiian-English Dictionary*. University of Hawaii Press.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Michal Richter, Pavel Straňák, and Alexandr Rosen. 2012. Korektor – a system for contextual spell-checking and diacritics completion. In *Proceedings of COLING 2012: Posters*, pages 1019–1028, Mumbai, India. The COLING 2012 Organizing Committee.

Tim Schlippe, ThuyLinh Nguyen, and Stephan Vogel. 2008. Diacritization as a machine translation and as a sequence labeling problem. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Student Research Workshop*, pages 270–278, Waikiki, USA. Association for Machine Translation in the Americas.

Brendan Shillingford and Oiwi Parker Jones. 2018. Recovering missing characters in old Hawaiian writing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4929–4934, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Eseta Magaui Tualaulelei, Fepuleai Lasei John Mayer, and Galumalemana A Hunkin. 2015. Diacritical marks and the Samoan language. *The Contemporary Pacific,*, Volume 27(Number 1).

Sunny Walker, Winston Wu, Bruce Torres Fischer, and Larry Kimura. 2025. Kuene: A web platform for facilitating Hawaiian word neologism. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 182–187, Honolulu, Hawaii, USA. Association for Computational Linguistics.

William H Wilson. 1981. Developing a standardized Hawaiian orthography. *Pacific Studies*, 4:19–19.

Annette Kuuipolani Kanahele Wong. 2020. *Mai Pukaiki Kula Maniania a Puuwai Aloha o ka Ohana*. University of Hawaii Press.

Ipo Wong and Kahea Faria. 2021. Ka leo o na kupa: Hawaiian language revitalization efforts by native (L1) speakers. Timestamp: 10:00.

Kuuipolani Wong, Keao NeSmith, William O'Grady, Bradley McDonnell, and Kainehe Chun-Lum. 2025. Research study of Olelo Niihau.

Winston Wu and David Yarowsky. 2021. On pronunciations in Wiktionary: Extraction and experiments on multilingual syllabification and stress prediction. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 68–74, Online (Virtual Mode). INCOMA Ltd.

David Yarowsky. 1994. A comparison of corpus-based techniques for restoring accents in Spanish and French text. In *Second Workshop on Very Large Corpora*, pages 19–32.

A Sample Model Predictions

Table 6 shows sample model output for two sentences in the test set.

Original	Ae. A, i keia manawa, a, o wai ka inoa o ke mele a oukou e himeni ana?
Gold	ʻAe. A, i kēia manawa, a, ʻo wai ka inoa o ke mele a ʻoukou e hīmeni ana?
Unigram	ʻAe. A, i kēia manawa, a, ʻo wai ka inoa ʻo ke mele a ʻoukou e hīmeni ʻana?
Bigram	ʻAe. A, i kēia manawa, a, ʻo wai ka inoa o ke mele a ʻoukou e hīmeni ʻana?
Trigram	ʻAe. A, i kēia manawa, a, ʻo wai ka inoa o ke mele a ʻoukou e hīmeni ana?
Unigram (Corpus)	Ae. A, i kēia manawa, ʻa‘, ʻo wai ka inoa ʻo ke mele ʻa‘ ʻoukou ʻē hīmeni ʻana?
Bigram (Corpus)	Ae. A, i kēia manawa, ʻa‘, ʻo wai ka inoa ʻo ke mele ʻa‘ ʻoukou ʻē hīmeni ʻana?
LSTM	ʻAe. A, i kēia manawa, a, ʻo wai ka inoa o ke mele a ʻoukou e hīmeni ana?
Transformer	ʻAe. A, i kēia manawa, a, ʻo wai ka inoa o ke mele a ʻoukou e hīmeni ana?
LLM	Ae, a, i keʻiā manawa, a, o wai ka ʻinō o ke mele a oukou e himenī ana?
LLM Finetuned	ʻAe. A, i kēia manawa, a, ʻo wai ka inoa o ke mele a ʻoukou e hīmeni ana?
Original	akoakoa mai hookahi hale pule a hoike aku a hoike mai
Gold	ʻākoakoa mai hoʻokahi hale pule a hōʻike aku a hōʻike mai
Unigram	ʻākoakoa mai hoʻokahi hale pule a hōʻike aku a hōʻike mai
Bigram	ʻākoakoa mai hoʻokahi hale pule a hōʻike aku a hōʻike mai
Unigram (Corpus)	ʻākoakoa mai hookahi hale pule ʻa‘ hōʻike aku ʻa‘ hōʻike mai
Bigram (Corpus)	ʻākoakoa mai hookahi hale pule ʻa‘ hōʻike aku ā hōʻike mai
LSTM	ʻakoʻakoʻakoʻa mai hoʻokahi hale pule a hōʻike maii hōʻike
Transformer	ʻākoʻakoʻa mai hoʻokahi hale pule a hōʻike aku a hōʻike mai
LLM	ʻAkoakoa mai hookahi hāle pūle a hoike aku a hoike mai
LLM Finetuned	ʻākoakoa mai hoʻokahi hale pule a hōʻike aku a hōʻike mai

Table 6: Sample predictions of each approach on the orthography conversion task. Because the n-gram models performed similarly, not all of them are shown.