Kuene: A Web Platform for Facilitating Hawaiian Word Neologism

Anonymous ACL submission

Abstract

This paper presents Kuene, a web-based collaborative dictionary editing platform designed to facilitate the creation and publication of Hawaiian neologisms by the Hawaiian Lexicon Committee. Through Kuene, the Committee can create, edit, and refine new dictionary entries with a multi-round approval process, ensuring accuracy and consistency. The platform's technical features enable flexible access control, fine-grained approval states, and support for multimedia content and AI-assisted orthography modernization. Just in the past two months, Kuene has enabled the publication of over 400 new Hawaiian words. By streamlining the dictionary editing process, Kuene aims to alleviate the scarcity of modern Hawaiian words and facilitate the revitalization efforts of the Hawaiian language.

1 Introduction

Hawaiian is a critically endangered language in the Austronesian language family, spoken in the state of Hawaii, USA. Through most of the 1900s, Hawaiian was banned in schools, leading to a sharp decline in usage and a generation with nearly no native speakers. Only in the past 40 years have there been active efforts to revitalize the language through educational initiatives such as immersion schools, leading to a resurgence of usage. One of the many hindrances to the active use of Hawaiian in daily life today is the lack of words for many modern concepts. To remedy this issue, the Hawaiian Lexicon Committee, Kōmike Hua'ōlelo, was formed in 1987 for the purpose of creating new words in the language. The Committee is composed of native Hawaiian speakers who meet regularly to discuss and create new words. As a result of their meetings, the Committee has published Māmaka Kaiao (Kōmike Hua'ōlelo, 2003), a dictionary of modern Hawaiian words, which has been updated several times since. This dictionary, along with oth-

ers (Pukui et al., 1976; Andrews, 1865; Pukui and Elbert, 1986), have been instrumental for students and learners of Hawaiian. However, due to several factors including the COVID pandemic, the Committee has not met in several years, and progress on updating Māmaka Kaiao with new words has stalled until very recently.

In this paper, we present Kuene, an online collaborative dictionary editing and publishing platform that facilitates the process of creating and publishing neologisms by the Hawaiian Lexicon Committee. Using Kuene, the Committee can propose new words and definitions. Then, other Committee members can review proposed entries, making edits as needed. Several rounds of approvals by different members can be completed through Kuene to ensure the accuracy of the new words, their translations, parts of speech, example usages, and other information associated with the new entry. After a final editorial review, a word can be seamlessly published using a one-click export to a public Hawaiian dictionary website.

Kuene sports several technical features that facilitate the neologism process. User accounts with different permissions can limit access to users with different roles, e.g. one member responsible for creating the dictionary entry, or an editor responsible for proofreading for typos. An entry's headword and definition can have different approval states, allowing for finer distribution of effort when approving a new entry, particularly for headwords that have previously approved definitions. Users may also post internal comments for in-context asynchronous discussion about entries. Kuene takes advantage of the web-based medium to support embedding of media such as photos, audio, video, and taxonomic tagging to further add context to dictionary entries, enhancing comprehension for new and multimodal learners of Hawaiian. Furthermore, Kuene supports efficient checking for duplicate entries and existing related entries and integration

with AI tools for NLP-assisted modernization of Hawaiian's 19th century printing press orthography.

Just in the past two months, the Kuene platform has been used to publish over 400 new Hawaiian words, and it is also being used to develop a dictionary for legal Hawaiian terms and a monolingual (Hawaiian-Hawaiian) dictionary. With Kuene, we envision a considerably shorter lead time from the proposal to the publication of new words by the Hawaiian Lexicon Committee, which will greatly alleviate the lack of missing words in Hawaiian as well as support the language revitalization efforts of this critically endangered language.

2 Related Work

To our knowledge, there are no existing software designed specifically to aid the revitalization of an endangered language through the creation and publication of new words. Two related areas are language documentation and conlang creation.

In terms of language documentation, software such as FieldWorks Language Explorer[1] and WELT (Ulinski et al., 2014) have been developed for field linguists to elicit and document words in a language. These tools have a number of features that are useful in language documentation, but they are not designed for the creation of new words. Furthermore, these complex tools are designed for trained linguists, while several Hawaiian Lexicon Committee members are elders with little technology experience. In addition, due to the potential lack of internet in documenting endangered languages, language documentation tools are often installed onto a computer, as opposed to Kuene's web-based interface. Some tools such as Linguistic Field Data Management and Analysis System (LiFE) (Singh et al., 2022) and Glam (Gessler, 2022) integrate NLP into the language documentation process, but these systems are also designed primarily for use by field linguists rather than the speakers of the language.

Conlangs (constructed languages) undergo a similar process as Hawaiian, where one goal of a conlang's creator is to expand the language's vocabulary. However, conlang vocabulary is often limited to the creator's needs, and the development of a conlang's vocabulary is often manageable with a simple excel spreadsheet. Some specialized software exists for keeping track of a conlang's vocabulary and

grammar, such as PloyGlot[2]; language documentation tools mentioned above can also serve this purpose. These kinds of software include features not needed by the Hawaiian Lexicon Committee, such as encoding phonological rules. Furthermore, creating conlangs is often a one-person affair, and as such, conlang software often do not support collaborative editing.

In the NLP literature, there is a wealth of research in detecting neologisms (e.g. Cartier, 2017; Breen et al., 2018; McCrae, 2019; Ryskina et al., 2020) and computationally constructing neologisms (e.g. Özbal and Strapparava, 2012; Das and Ghosh, 2017; Mizrahi et al., 2020). However, these tasks are not the current focus of Kuene, which aims to facilitate the *human* process of creating neologisms.

3 System Description

We describe the components of the Kuene platform and how it facilitates the work of the Hawaiian Lexicon Committee. A screenshot of the entry editing page for the word *pena waha* (lipstick) is shown in Figure 1. The interface can be localized in Hawaiian or English and easily supports extension to other languages.

Entry. The main unit of data in Kuene is the entry. A user can enter a new entry in either the Hawaiian to English direction or the English to Hawaiian direction, which we refer to as the Headword entry and Definition entry, respectively. Both directions will ultimately need to be entered, but Kuene supports automatically creating the opposite direction entry and linking to existing entries. The main components of a Headword entry are: the Hawaiian word, segmented syllables (to aid in pronunciation), approval state, and potential links to other headwords. The Definition entry is more complex, containing the English definitions, the part of speech, and several other attributes described below. Kuene currently supports 14 Hawaiian parts of speech defined in Kamanā and Wilson (2012). Kuene is also inherently multilingual, supporting headwords and definitions in English and Hawaiian, as well as other languages including Latin, French, Maori, Samoan, and Tahitian.

Approval. A dictionary entry must undergo numerous checks for quality by different people before being published in the dictionary. To support finer division of labor, the headword entry and definition
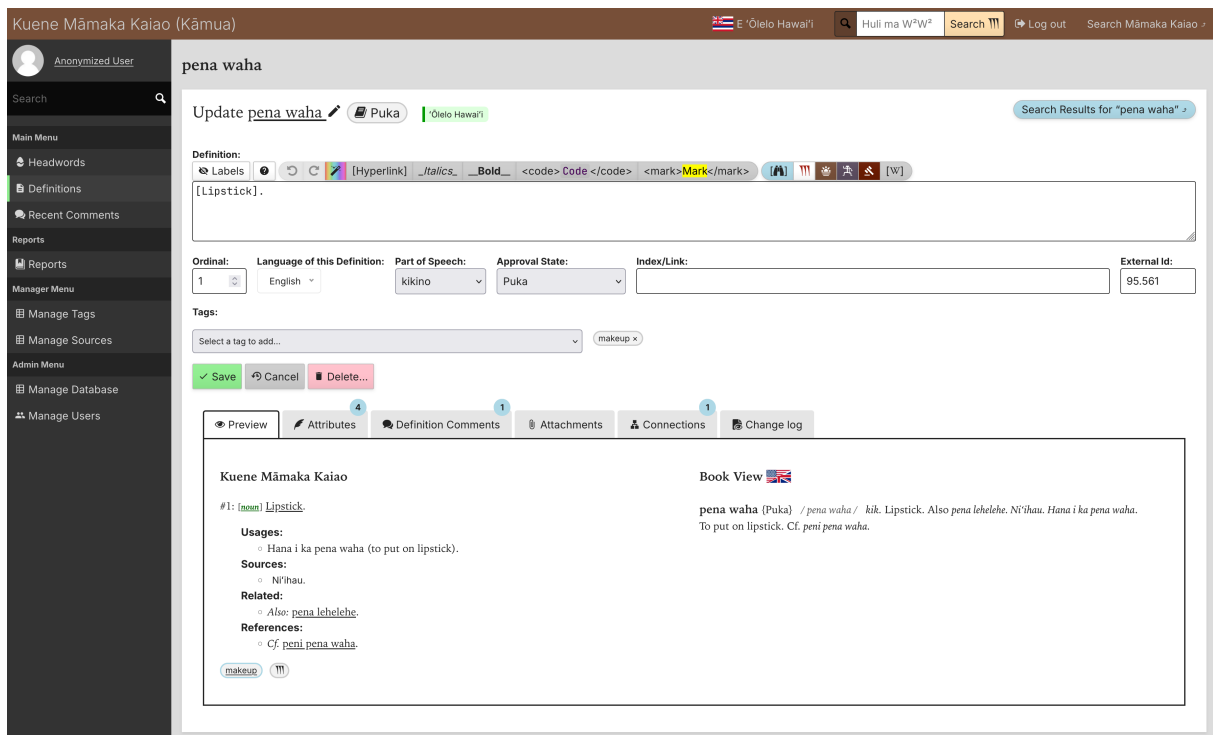
Figure 1: A screenshot of Kuene on the editing page for the word *pena waha* (lipstick), one of many words which was added to the Hawaiian dictionary in the past year. In Appendix A, Figure 3 shows the same interface for a user with lower access privileges.

entry can be separately approved. This is particularly useful for cases where definitions (English words) have already been approved (i.e. there is a demonstrated need for a Hawaiian neologism), but the Hawaiian word is still being considered by the committee. Kuene supports a variable number of approval states as designated by the Committee's needs, ranging from *introduced* to *needs supplemental attributes* to *published*.

Attributes. Entry attributes provide details about an entry that aid in comprehension of the word, including example sentences, sources where this word was used or found, related words, references, and tags to indicate the topical categories of an entry. Kuene has built-in functionality to format each attribute with appropriate HTML styling when published to the dictionary website or output to a printed version, and also includes a helpful feature to automatically add missing diacritics with AI integration (described below). In Appendix A, Figure 2 presents an interface (localized to Hawaiian) for editing the available tags.

Comments. Because users of Kuene live in various parts of the Hawaiian Islands, it is important that users can communicate about an entry without having to travel to the same place. The Comments section allows for in-context asynchronous discussions about an entry, facilitating the Committee's work.

Attachments. The Attachments section allows users to upload images, audio files, and video clips that can be displayed on the dictionary website or to support discussion about an entry.

Changelog. For accountability of online management, Kuene supports an in-context change log so users can view a timeline of entry modifications, preserving historical data.

Reports. Kuene can generate reports that summarize the current state of progress. Currently, Kuene supports generating reports for Approval States, Attachments, Attributes, Recent Changes, Duplicate Headwords, and Tags. This allows users, for example, to quickly list all entries that are at a specified approval state, with links to edit those entries.

Publication. Once an entry has obtained full approval, automated checks can be performed for existing duplicate or related entries, which can be manually fixed by editing the entry through Kuene. After all necessary edits have been made, Kuene

3

can export the entry to an existing online Hawaiian dictionary[3] for use by the general public. If additional edits need to be made, the user can perform the edits through Kuene and then republish the entry.

3.1 Orthography Modernization

When Protestant missionaries first arrived in Hawaii around 1820, they introduced an orthography using Latin letters to the previously unwritten Hawaiian language. The use of the ʻokina (ʻ) to represent glottal stops, and the kahakō (macrons) for long vowels, was not standardized until about 100 years later. Today, text with the ʻokina and kahakō diacritics is particularly helpful for new Hawaiian language learners who are not able to easily discern between words that are spelled the same without diacritics by context alone.

Kuene also supports integration with AI tools for NLP-assisted modernization of the 19th century missionary orthography. When viewing selected example sentences from old sources written in the old orthography, users can have AI systems perform a first pass at modernizing the orthography of the entry. On the backend, this is implemented with an API call to OleloTools[4], which uses a custom ChatGPT prompt. Preliminary experiments show that this method is competitive with sequence-to-sequence Transformer translation models. After orthography modernization by an NLP model, the user can make necessary corrections before saving their edits to the dictionary entry. There is also a feedback mechanism where users can indicate that their corrections will be saved for future retraining of the orthography modernization model.

4 Conclusion

We presented Kuene, an online collaborative dictionary platform that facilitates the work of the Hawaiian Lexicon Committee in coining Hawaiian neologisms. Kuene supports all the steps of coining new Hawaiian words, from the creation to publication in a public online Hawaiian dictionary. Kuene has already seen major successes, with over 400 words published through Kuene in the last two months. The design of Kuene is very modular and extensible, making it relatively easy to produce different dictionaries for different purposes or even different languages. Kuene is also currently being used to develop a dictionary for historical legal Hawaiian terms and a monolingual (Hawaiian-Hawaiian) dictionary.

Because Kuene supports multiple dictionary sources, a long-term goal is to develop a unified corpus of entries drawn from various Hawaiian dictionaries available today. We also plan to expand Kuene to incorporate more NLP methods and techniques, including improvements to the orthography modernization, and tools that can automatically generate neologism, e.g. Özbal and Strapparava (2012), which can potentially lessen the cognitive load of the Committee members. Thanks to the efforts of the Hawaiian Lexicon Committee, the new words added to the Hawaiian lexicon through Kuene are contributing significantly to the revitalization of Hawaiian and promoting wider use of the Hawaiian language in daily life.

Limitations

Our paper presents a web platform to assist the Hawaiian Lexicon Committee in creating neologisms for Hawaiian, a critically endangered and historically marginalized language. This type of system may not be applicable to all under-resourced languages in need of new vocabulary, especially if there is no committee or governing body responsible for introducing new words to the language's lexicon. We also recognize that there may be differing opinions about the best way to facilitate the neologism process. For example, Hornsby and Quentel (2013) describes conflicts of authenticity of neologisms from different sources in Breton, a Celtic language spoken in Brittany in modern-day France. In this paper, we have taken a pragmatic approach, focusing on creating a user-friendly and functional platform that meets the needs of the Hawaiian Lexicon Committee.

References

Lorrin Andrews. 1865. *A dictionary of the Hawaiian language, to which is appended an English-Hawaiian vocabulary and a chronological table of remarkable events*. Dalcassian Publishing Company.

James Breen, Timothy Baldwin, and Francis Bond. 2018. The company they keep: Extracting Japanese neologisms using language patterns. In *Proceedings of the 9th Global Wordnet Conference*, pages 163–171, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Emmanuel Cartier. 2017. Neoveille, a web platform for neologism tracking. In *Proceedings of the Software*

---

[3]https://anonymized
[4]https://olelotools.com/

Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 95–98, Valencia, Spain. Association for Computational Linguistics.

Kollol Das and Shaona Ghosh. 2017. Neuramanteau: A neural network ensemble model for lexical blends. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 576–583, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Michael Hornsby and Gilles Quentel. 2013. Contested varieties and competing authenticities: neologisms in revitalized breton. *International Journal of the Sociology of Language*, 2013(223):71–86.

Kauanoe Kamanā and William H Wilson. 2012. *Nā Kai ʻEwalu*. Hale Kuamoʻo.

Kōmike Huaʻōlelo. 2003. *Māmaka Kaiao: A Modern Hawaiian Vocabulary*. University of Hawaiʻi Press.

John Philip McCrae. 2019. Identification of adjective-noun neologisms using pretrained language models. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 135–141, Florence, Italy. Association for Computational Linguistics.

Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. 2020. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4918–4929, Online. Association for Computational Linguistics.

Gözde Özbal and Carlo Strapparava. 2012. A computational approach to the automation of creative naming. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 703–711, Jeju Island, Korea. Association for Computational Linguistics.

Mary Kawena Pukui and Samuel H Elbert. 1986. *Hawaiian dictionary: Hawaiian-English English-Hawaiian revised and enlarged edition*. University of Hawaii Press.

Mary Kawena Pukui, Samuel H Elbert, and Esther T Mookini. 1976. *Place Names of Hawaii: Revised and expanded edition*. University of Hawaii Press.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 367–376, New York, New York. Association for Computational Linguistics.

Siddharth Singh, Ritesh Kumar, Shyam Ratan, and Sonal Sinha. 2022. Towards a unified tool for the management of data and technologies in field linguistics and computational linguistics - LiFE. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 90–94, Marseille, France. European Language Resources Association.

Morgan Ulinski, Anusha Balakrishnan, Daniel Bauer, Bob Coyne, Julia Hirschberg, and Owen Rambow. 2014. Documenting endangered languages with the WordsEye linguistics tool. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 6–14, Baltimore, Maryland, USA. Association for Computational Linguistics.

A Additional Screenshots

Figure 2 and Figure 3 present additional screenshots showing the powerful functionality of Kuene.
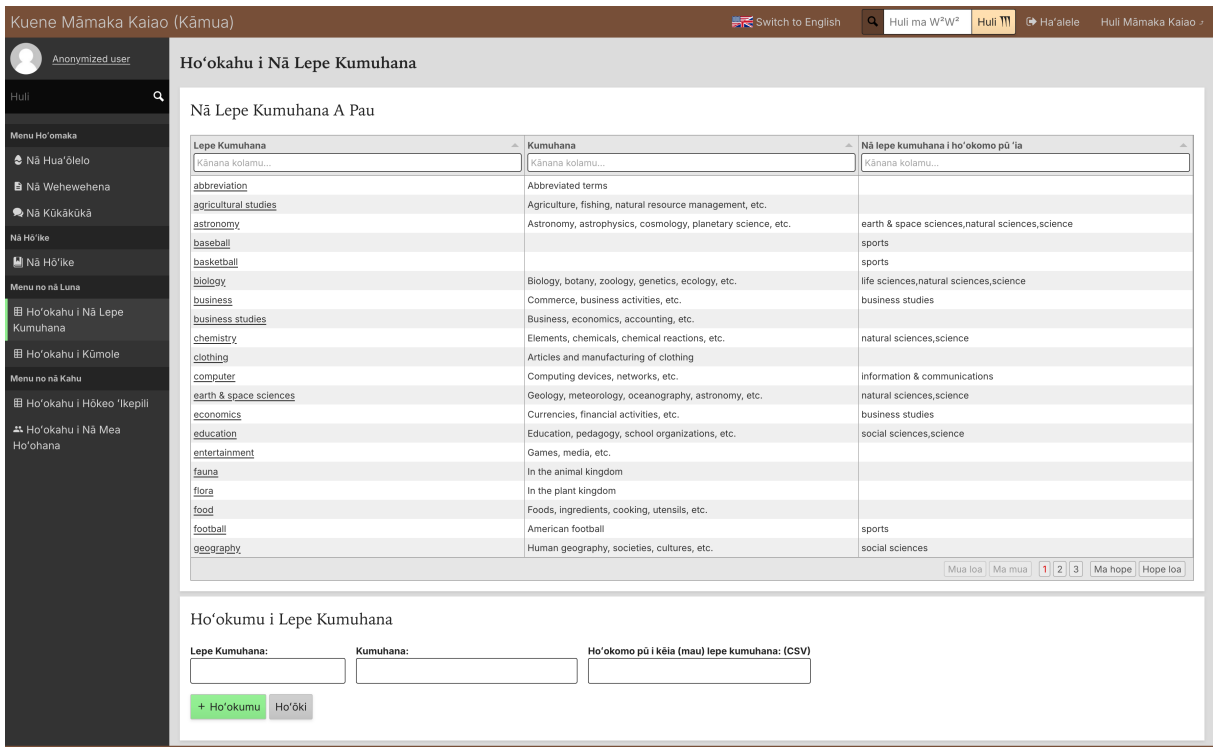
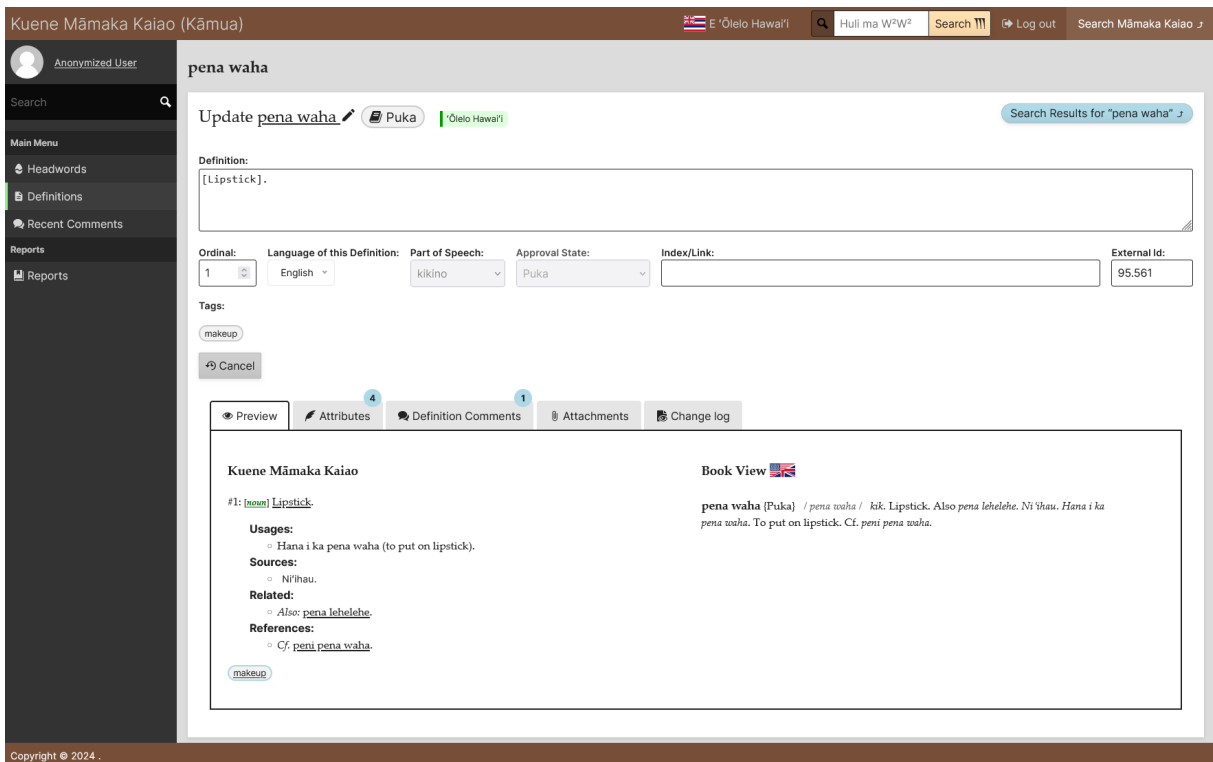Figure 2: Kuene supports localized UI and assigning tags for categorizing entries.

Figure 3: Editing the word *pena waha* from the perspective of a user with lower access privileges and an English interface.