Developing a Hawaiian Corpus Toolkit for Data-Driven Language Learning

Joseph Winkie
University of Hawai‘i at Hilo
jwinkie@hawaii.edu

Michol Malia Miller
University of Hawai‘i at Mānoa
michol@hawaii.edu

Winston Wu
University of Hawai‘i at Hilo
wswu@hawaii.edu

Abstract

This paper presents the development of an online multimodal corpus toolkit designed for data-driven language learning in Hawaiian. The toolkit supports corpus linguistics analyses including concordance/KWIC (Key Word In Context) searches, frequency analysis, collocation analyses, and complex queries with n-grams and regex pattern matching. Specifically designed for educators, students, and parents within the Hawaiian community, this easy-to-use tool facilitates a data-driven language learning process by enabling users to explore authentic language data, identify patterns, and develop deeper understanding of Hawaiian language structures through computational methods. By integrating corpus-based approaches into language education, this toolkit contributes significantly to preserving and promoting Hawaiian language learning and supports the broader community's efforts in language revitalization.

1 Introduction

The lack of language teaching materials is a major challenge for Indigenous language revitalization (ILR). In these low-resource language communities, language teachers frequently design and produce their own teaching materials for use in language classes. Oftentimes, teachers and learners do not have sufficient opportunities to interact with fluent speakers who can provide the authentic, naturalistic oral language input necessary for developing second language speaking proficiency. Without opportunities for such authentic language input, teachers and learners must turn to their community's language archives. The creation and adaptation of language corpora and corpus-based tools to support data-driven language learning offers a promising way forward to address these challenges in Indigenous language education.

Data-driven learning (DDL, Johns, 1991) is a pedagogical approach in which a collection of texts, or corpus, is used to explore language patterns in the classroom using computerized tools. Extensive studies on the use of corpus linguistics and DDL for English language learning and materials development (O'Keeffe and McCarthy, 2022; O'Keeffe et al., 2007; Sinclair, 2004) have reported that DDL benefits language learners' acquisition of vocabulary and lexicogrammar (Boulton and Cobb, 2017) and fosters learner autonomy (Charles, 2022). One aspect particularly important for ILR is that DDL can provide authentic language input for teachers and learners in the absence of fluent speakers, giving them the ability to access examples of specific grammatical structures and vocabulary (Römer, 2011).

‘Ōlelo Hawai‘i (Hawaiian) is a critically endangered Polynesian language. After almost losing a generation of native speakers due to the suppression of Hawaiian in public schools, community interest in revitalizing and re-normalizing Hawaiian has grown over the last 40 years. Current efforts to revitalize the language are mostly in language classes in schools (including immersion schools, public K-12 schools, and universities) and through community efforts to speak Hawaiian at home. However, because suppression of the language resulted in gaps in intergenerational language transmission, most teachers, though fluent, are not native speakers. DDL has the potential to greatly improve the learning experience for Hawaiian learners, by allowing them to learn from a corpus of natural Hawaiian speech produced by native speakers.

For ‘ōlelo Hawai‘i, corpus-driven approaches to language *research* are not new. Hawkins (2003) utilized corpus-based methods to investigate the distribution and pragmatic functions of the verbal particle *ana* in 18th and 19th century texts. Baker (2012) analyzed the use of genitive subject class selection in nominalizations and relative clauses in traditional stories sourced from Hawaiian language newspapers. More recently, Hosoda (2019)

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 167–176

applied corpus analyses to catalog and explore the usage of Hawaiian morphemes in dictionary data to support Hawaiian language information retrieval. Brockway (2021) generated frequency-based word lists for the semantic frame of *ʻāina* (land) from a small corpus of selected transcripts of the Ka Leo Hawaiʻi radio program. The Combined Hawaiian Dictionary (Trussel, 2022) hosts an extensive catalog of Hawaiian concordances and topical vocabulary lists compiled using corpus analysis from key reference materials, including the most frequently utilized Hawaiian dictionaries and the Hawaiian Bible. However, the use of DDL as a form of *classroom pedagogy* and the use of corpora to create learning materials for low-resource Indigenous languages such as Hawaiian remains underexplored.

This paper describes the development of a web-based multimodal corpus toolkit that offers teachers and learners of Hawaiian the ability to engage in DDL in the classroom through the use of corpus linguistics analyses including concordance analysis, the generation of frequency-based word lists, and displaying collocations. We describe the features and system architecture of our toolkit, followed by use cases, initial interviews with Hawaiian language teachers, and future plans.

2 Related Work

Some recent examples of the application of corpus-based approaches for Indigenous language revitalization include community-based corpus compilation work with nêhiyawêwin (Plains Cree) (Teodorescu et al., 2022), as well as with Stoney and Dene Sųłiné (Rice and Thunder, 2017); Kven in Norway (Lane et al., 2022); and Cherokee (Frey, 2018). These studies largely focus on the process of corpus creation, rather than a toolkit to interact with the corpus.

On the corpus linguistics tools side, there are a handful of existing systems with goals similar to our own, but are limited in their applicability to Hawaiian. Sketch Engine (Kilgarriff et al., 2014) is an online, proprietary corpus tool for searching corpora. It can analyze collocations, concordances, wordlists, word trends, and several other features. It contains built-in corpora for 100+ languages, although Hawaiian is not one of them. The major disadvantage of Sketch Engine that it requires a monthly subscription, which is often prohibitive to members of low-resource language communities. A free version, NoSketch Engine, remedies the cost issue but removes many features that are useful for supporting language education and revitalization. The web interface to the Corpus of Contemporary American English (COCA Davies, 2009) also supports similar features of collocations, concordances, and word frequency, but it only supports text in English, specifically English published online in the last 20-30 years, while our focus is on Hawaiian.

Perhaps most similar to our work is AntConc (Anthony, 2005), a freeware corpus analysis software that supports concordance, word and keyword frequency, and wildcard searching. The disadvantage of AntConc is that it does not come with its own corpora; the user needs to import data into the program. This is prohibitive for students and teachers who may have little computational background, and also tedious because the user must collect a corpus first. In contrast to these, our system is fully online, requiring no downloading, and comes preloaded with Hawaiian corpus data, ready for the user to query.

Other examples of using corpus-based tooling for linguistic studies and education include Xu et al. (2012), who created a database containing sentences selected by linguistics experts and linguistic facts covered in an authoritative Chinese Reference Grammar. Coole et al. (2020) created a database specifically designed to scale well while efficiently handling queries created by tools such as Key Word In Context (KWIC) and collocation search, and showed that LexiDB improved performance over other NoSQL databases such as MongoDB and Cassandra. With the relatively small dataset available for Hawaiian, we were able to get sufficient performance out of PostgreSQL, due to its powerful ability to sort, filter, combine, and manipulate data, which allowed us to perform most data retrieval operations in a single query.

3 Data

The corpus for this study was compiled from a widely used set of spoken Hawaiian language materials representing the register of conversational Hawaiian. The data consists of audio recordings and transcripts from the *Ka Leo Hawaiʻi* radio program, hosted in the Kaniʻāina digital repository (Kimura, 2025). The *Ka Leo Hawaiʻi* collection, which ran from 1972-1988, consisted of a series of interviews conducted by Larry Kimura with native Hawaiian speakers, in addition to listener calls and musical performances. This collection was chosen

due to its extensive use for Hawaiian language education by teachers and learners. Of the 417 *Ka Leo Hawai'i* episode recordings available online, only 45 episodes are provided with corresponding transcripts. These recordings were segmented to create time-aligned annotations, and transcripts were manually written. These transcripts and corresponding audio recordings make up the present corpus.

Because the *Ka Leo Hawai'i* collection is a finite set of recordings, the corpus can best be described as an opportunistic corpus. Opportunistic corpora consist of the amount of data possible to gather due to limitations such as lack of funding or low numbers of speakers available for documentation (or in this case, the end of the radio program in 1988), and are more common in the case of endangered languages (McEnery and Hardie, 2012). As a result, during the design process, it was not possible to implement a rigorous sampling scheme to ensure representativeness, balance, and mitigate issues of skew for this corpus.

3.1 Data Preparation

For efficient querying, we store the text in a database. To prepare the text for ingestion into our database, we first preprocess the text bylowercasing all letters and preserving digits, the ʻokina, and vowels with diacritics (ā, ē, ī, ō, ū). Any character outside this set (such as punctuation or whitespace) is treated as a delimiter to split the text. This approach ensures that the reverse index handles Hawaiian-specific orthography correctly while keeping the database entries consistent and case-insensitive.

3.2 Corpus Statistics

The dataset contains 555,707 total word tokens and 7,824 unique word forms. Because the pretokenization logic focuses on string-matching rather than morphological analysis, we index these as raw surface forms rather than distinct lemmas. This corpus size is sufficient for a functional reverse index; 7,824 unique words generally cover the vocabulary needs of an intermediate proficiency level, capturing the vast majority of terms used in standard and educational Hawaiian contexts.

4 System Features

Our corpus toolkit is designed to be accessed using a web browser and currently supports four main features that are useful for both teachers and learners of Hawaiian. A screenshot of the homepage is shown in Figure 1.

4.1 Corpus Linguistics

Our platform supports several common corpus linguistics analyses, described below.

4.1.1 Concordance/KWIC

A concordance is a list of every occurrence of a word along with its surrounding context. Also known as keyword in context (KWIC), a concordance allows the user to find and compare usages of a word. In the past, concordances were manually created for important works such as the Bible, but with modern technology, concordances are easily created computationally. For a language learner, concordances allow the student to see many naturalistic examples of a word that they have just learned; the ability to visualize these examples with a concordancer can support inductive and analytical language learning (Vyatkina, 2020) by encouraging learners' pattern recognition (Boulton and Cobb, 2017). For teachers, consulting a corpus can supplement teaching materials with numerous example sentences to provide learners with linguistic evidence in the form of concordance lines (Tsui, 2004).

4.1.2 Frequency Lists

Frequency lists show the number of occurrences of a word or phrase. For language learners and teachers, frequency lists are important for determining which words should be prioritized in the learning process. A word that occurs frequently in a language should be memorized in the early stages of learning, while a relatively rarer word can be looked up when encountered. Frequency lists can also be generated for a subset of a corpus consisting of a single text or a selection of texts; it would be helpful for learners who are trying to read a new piece of literature to first learn the most common words within. For majority languages like English, there are several such lists for language learning derived from corpora, including the New General Service List (Brezina and Gablasova, 2015) and the New Academic Word List (Gardner and Davies, 2014).

4.1.3 Collocations

Collocations are made up of words that co-occur more frequently than would be expected by chance. Proficiency in using collocations has long been seen as important for language learners (Palmer, 1933).

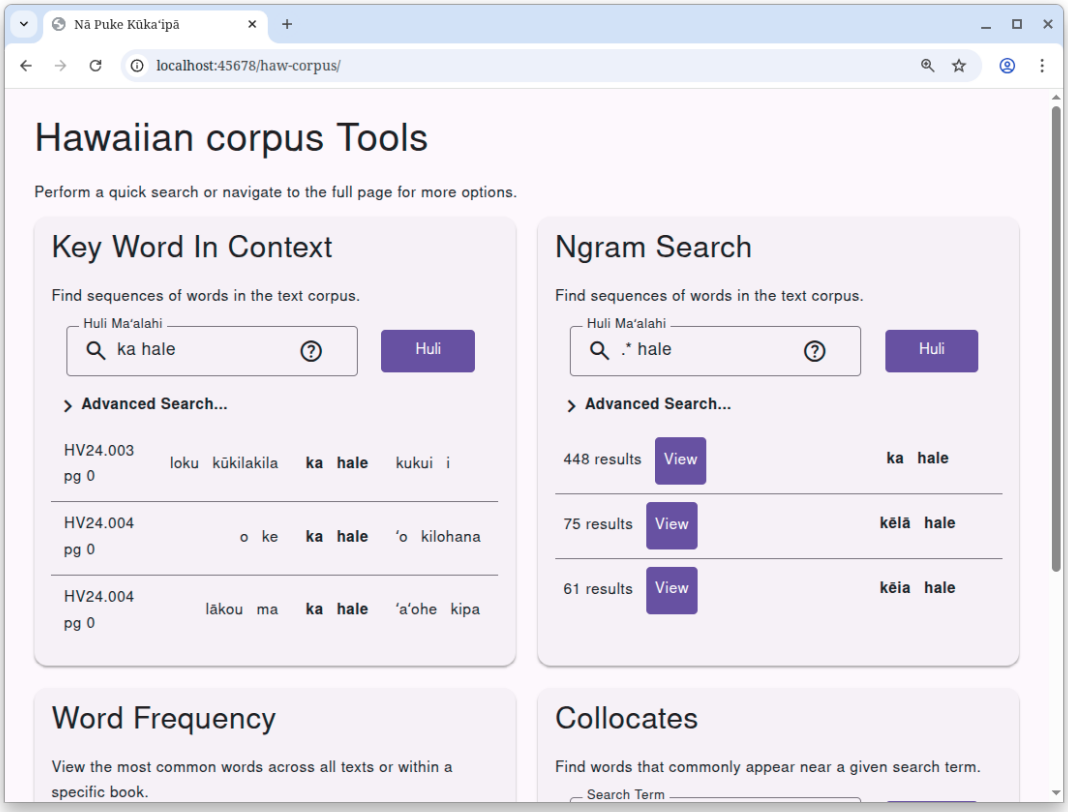

Figure 1: Screenshot of the Homepage of the Hawaiian Language Corpus Toolkit

Similar to concordances and frequency lists, identifying a word's collocations allows learners to see the most frequent usages of a word in different, potentially idiomatic, contexts.

4.1.4 N-gram and Regex Search

The above three analyses are useful for single words, but can be made even more powerful by allowing the user to search for n-grams (multiple words) or employ searches involving regular expressions (regex). This functionality addresses a key limitation of traditional corpus tools that only allow single-word queries, enabling users to explore linguistic patterns across word boundaries or words with the same morpheme.

For example, *wale nō* (only, just) is a common phrase that a student would like to search, but searching for *wale* (slime, mucus) would include other usages. Regular expressions allows the user to search for complex queries to fit their needs. For example, searching for *hoʻo* .* would find all words starting with the causative *hoʻo* prefix (e.g. *hoʻohiki, hoʻonani*). Our implementation of regex search also supports searching with multiple words. For example, the search .* *ʻana* would find all verbs that have been turned into gerunds, and *ʻaʻole* .* *i* would find negative past tense phrases. This level of search capability transforms the corpus from a static reference into an active discovery tool, empowering learners to formulate and test linguistic hypotheses about Hawaiian grammar, vocabulary, and usage patterns drawn from real-world texts.

4.2 Other Features

4.2.1 Audio

The majority of our corpus is sourced from audio-based sources such as radio interviews, providing language learners with the opportunity to hear natural speech from native speakers of the language. Users are able to navigate through the corpus tool to either the Key Word In Context page or the full page text, and play these recordings line by line as they read the text.

Most spoken corpora available for data-driven learning are mono-modal and text-based, comprised of transcripts of audio recordings and do not contain sound files that would allow users to utilize search results for listening, pronunciation, and speaking practice (Crawford, 2022; Knight and Adolphs, 2021). With the combination of data from two modes, i.e., text and audio data, our corpus

can be classified as a multi-modal corpus. The inclusion of audio excerpts with each search query makes the development of this toolkit a unique contribution to the field of spoken corpus linguistics. The implementation of our multimodal corpus with teachers and learners of Hawaiian may provide new insights regarding the use of multimodal corpora for the acquisition of speaking skills.

4.2.2 Filtering

All of the lookup tools on our corpus tool suite allow for filtering by metadata. Users can filter by certain aspects of the data, such as which collection the text is from, where the speaker was born, if there is an audio track to listen to, who is speaking, and various other metadata. This information may be useful in certain cases, for example, if a student wants to listen to a specific speaker, or to mimic the accent of a native speaker from the same island as they were born.

4.2.3 Persistent Links

The page URL for a search contains all the data required for the backend to reproduce any result set. This enables teachers to share the exact search results with their students, enabling reproducibility. This was a highly requested feature by teachers.

4.3 Export Search Results

The search result pages of the corpus tool include an export utility allowing users to export a subset or all of their results to a CSV or TSV file, allowing for easy inclusion in spreadsheets or learning materials. This makes it easy for instructors and language learners to use the corpus as a resource to aid in the creation of additional educational materials, or to save the search results for future study.

Screenshots of several of these features are shown in Figures 2 to 4.

5 System Architecture

Our software stack consists of a web-server written in Go that serves a front-end and fetches data from a PostgreSQL database. This architecture is designed to leverage the power and efficiency of PostgreSQL and its query planner, while keeping the rest of the system simple. The source code for this server is intended to be freely available on GitHub once mature enough for release. Generally, operation starts with the user making an HTTPS web request to the server; the server then parses the request and builds

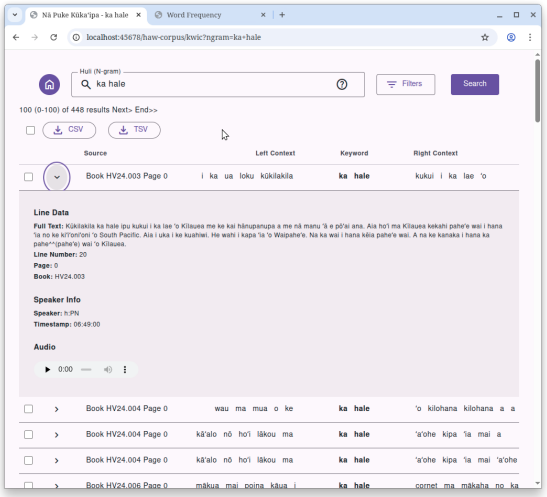

Figure 2: Screenshot of the KWIC result page for ”ka hale” in any book.

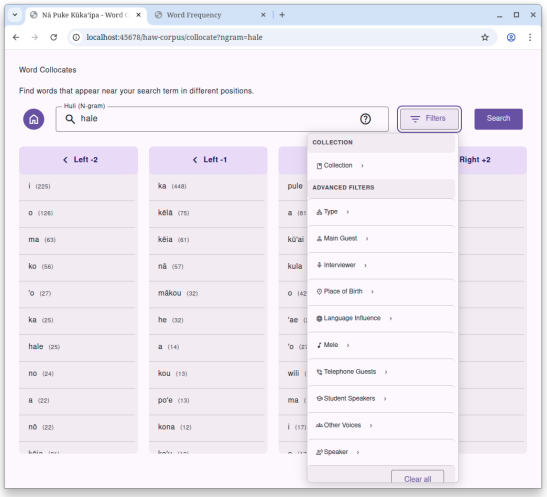

Figure 3: Screenshot of the collocates result page for ”ka hale”.

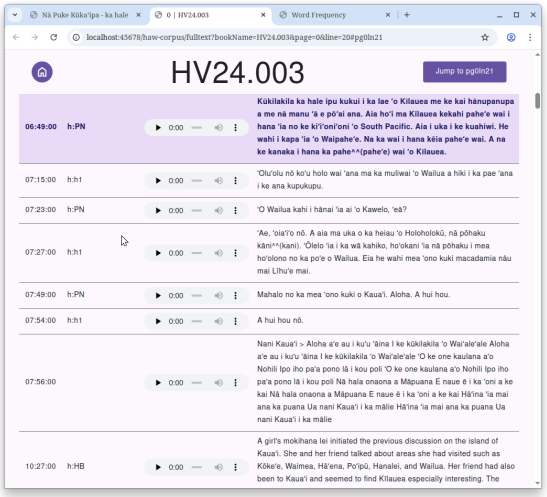

Figure 4: Screenshot of the Full-text page for a result

```
--- Generated SQL ---
WITH sequence_check AS (
    SELECT
        idx.book_name, idx.page, idx.line_number,
        idx.word, idx.word_num,
        LEAD(idx.word, 1) OVER (
            PARTITION BY idx.book_name
            ORDER BY idx.word_num
        ) as word1
    FROM index_data idx
    JOIN programs p ON idx.book_name = p.program_id
    WHERE (p.collection = $1)),
    filtered AS (
        SELECT book_name, page, line_number,
        word_num, word, word1 FROM sequence_check
        WHERE word ~ $2 AND word1 ~ $3
    )
    SELECT book_name, page, line_number, word_num,
        word, word1, COUNT(*) OVER() as total_count
    FROM filtered
    ORDER BY book_name, word_num LIMIT $4 OFFSET $5
--- Arguments ---
$1 = Ku i ka Manaleo
$2 = ^ka$
$3 = ^hale$
$4 = 100
$5 = 0
```

Figure 5: A generated SQL query for finding word sequences, with its runtime arguments.

a SQL query that handles sorting, filtering, and pagination. Finally, the server renders the data returned from the database into static HTML templates and serves them back to the user. Throughout this process, resource usage for both PostgreSQL and the web server are minimal.

5.1 Database

Fast searches are essential for a functional corpus toolkit. The backbone of our corpus toolkit is PostgreSQL and its ability to quickly retrieve batches of data. For efficient lookup, we store an inverted index, a data structure that facilitates the efficient retrieval of documents that contain a certain word. The inverted index is a table that maps each word in the corpus to the book, page, and word and line numbers where the word occurred. When a user searches for a word or phrase, we search each word in the inverted index partitioned by book, join on the book and author metadata for additional filtering, and then retrieve one ”page” of results. To achieve this, a large portion of the web server code is dedicated to assembling SQL queries such as the one in Figure 5. PostgreSQL also supports fast regular expression searches.

5.2 User Interface

The web server delivers a static and minimal, functional user interface (UI) designed for efficient corpus analysis by teachers and students. The homepage, depicted in Figure 1, presents the four primary analysis tools: Key Word in Context (KWIC), N-gram Search, Word Frequency, and Collocations. Each tool is presented in a distinct panel with simple input forms, allowing users to quickly initiate a query. Upon submission, the user is directed to a results page that displays the relevant data in a clean, legible list format, with the search terms bolded for easy identification.

A key design feature for promoting reproducibility of searches is the direct encoding of all search parameters into the page's URL. Every query component, including search terms, filters, and tool-specific options, is captured in a GET parameter. This architecture ensures that a URL is a persistent, shareable artifact that allows users to replicate the exact same analysis and view the identical result set simply by visiting the link. Consequently, teachers can easily share results links with their students, facilitating the learning process.

6 Use Cases

While the software has potential applications for linguistic research, its primary design objective is to serve as an educational tool for the Hawaiian language community. The interface and functionalities are optimized for students, educators, and parents engaged in language revitalization and education.

6.1 Language Learners

The platform is designed to facilitate data-driven learning, allowing students to move beyond rote memorization and engage with authentic language as it occurs in natural contexts. This approach fosters inductive learning, enabling students to discover linguistic patterns on their own. For example, a student who has just learned a new word or structure can use the Concordance/KWIC search to explore numerous examples of that word used in real-world conversation. Easy access to the audio clip associated with each concordance line also provides students with spoken language input, supporting the development of listening comprehension and accurate pronunciation.

6.2 Educators

Teachers can leverage the toolkit to create dynamic and evidence-based instructional materials. Rather than relying solely on textbook examples, an educator can query the corpus to find authentic sentences that illustrate a specific grammatical point or vocabulary theme. Additionally, the N-gram and Word

Frequency tools can help identify frequently used words and phrases to prioritize in lessons, ensuring that instruction is focused on high-utility language. An additional benefit of the toolkit is that users can perform corpus searches to test their assumptions and intuitions about language (Curry and Mark, 2024).

Furthermore, the shareable URLs are invaluable for assignments; a teacher can craft a specific query, send the link to students, and have the entire class analyze the exact same data set for an exercise or discussion. An additional function of the platform useful for teachers is the ability to export specific examples into a CSV or TSV file format, which they can then adapt into worksheets and handouts for the students.

6.3 Parents

Our toolkit also serves as a bridge for parents supporting their children's language journey, particularly when the parents might not be fluent speakers themselves. When a child needs help with homework or asks about a word, a parent can use the simple interface to search for it alongside the child. Reviewing the contextual examples together reinforces the child's learning and empowers the parent to be an active participant in their education.

6.4 Researchers

Although not the primary focus, the tool provides significant value for preliminary linguistic inquiry. Researchers can quickly perform exploratory analyses, such as identifying common collocations for a search term or examining word frequencies across different texts. The ability to easily share a direct link to a specific query result makes it a convenient tool for collaborating and citing specific examples from the corpus in research papers or presentations.

7 Community-Informed Corpus Development

Although recent meta-analyses support the effectiveness of DDL for language learning (Pérez-Paredes, 2022; Boulton and Cobb, 2017), relatively few teachers have integrated the use of corpora and DDL activities into language classrooms, due to a lack of teacher training opportunities and the complexity of learning to use corpus technologies (Ma et al., 2024). One issue is that large, well-known corpora are primarily used for linguistic research, and have not specifically been designed for classroom use by teachers and learners. In response, community-based, participatory research methods in partnership with teachers and learners have been proposed as a way to develop corpora that are both relevant and practical for pedagogical applications in contemporary language learning contexts (Curry and McEnery, 2025). Community-based research is a collaborative approach that centers the needs of language communities in language revitalization, and should focus on action-oriented research topics that are practical and relevant to each community (Rice, 2018). An important goal of community-based research is capacity-building, or training community members to engage in and eventually take over language research themselves (Czaykowska-Higgins, 2009). Following a participatory, community-based approach to corpus design allows developers to first gather the needs and concerns of teachers and learners, and then take action to address those needs when designing a pedagogical corpus. Training members of the Hawaiian language community to use the corpus for their own teaching and learning purposes will be an important step towards achieving the goal of capacity building in community-based research.

7.1 Needs Analysis and User Feedback

In the development and refinement of our Hawaiian corpus toolkit, we implemented a participatory approach to engage experienced Hawaiian teachers and their learners in the development process. This involves a multi-stage process that seeks to understand their needs and gather their feedback on our tool's ease of use. The first phase consisted of conducting interviews with experienced Hawaiian teachers to identify the needs for improving the oral language proficiency of students (especially regarding pronunciation, listening, and speaking development), as well as to collect information on current practices and challenges in utilizing the *Ka Leo Hawai'i* collection in Hawaiian language classes. Following the development of our toolkit, teachers were invited to use our toolkit and provide feedback for further improvements. The information collected in this phase will inform further improvements and changes to the toolkit.

We conducted initial interviews with two experienced Hawaiian language teachers, who have identified the following needs for Hawaiian learners. One teacher acknowledged that students need more spoken language input from a greater number of speakers than what they are exposed to in classes

with a single teacher and their classmates. Another teacher added that there are not enough opportunities for students to hear spoken Hawaiian in public and community spaces. Both teachers noted that, without language input from fluent speakers, learners tend to rely on translating their thoughts from English into Hawaiian, resulting in unnatural expressions that can lead to language change. One teacher also noted that some students experience insecurity when speaking, which can be a hindrance to developing their speaking skills; these students often wish to sound more like first language (L1) native speakers of Hawaiian. Another teacher said that, if learners cannot grow up around native speakers, at least they can listen to them in recordings.

When asked about current teaching practices using the *Ka Leo Hawai'i* collection, one teacher said they use excerpts for transcription activities to help develop learners' listening comprehension, as well as pronunciation activities where learners shadow or mimic L1 speakers' speech. For beginning learners, shorter excerpts with slower speech are used in lessons, while longer excerpts with a faster rate of speech are suitable for intermediate to advanced learners. This teacher also noted that using clips of authentic L1 speech in lessons helps expose students to variation in pronunciation, lexical, and grammatical structures across speakers.

After interacting with a preliminary version of our corpus tool, the teachers shared several positive reactions. One teacher highlighted the benefit of visualizing numerous examples at once using the KWIC/n-gram features. After looking through the concordance lines, the teacher discussed the possibility for learners to build their understanding of general patterns of language features, while also identifying exceptions to those patterns. The teacher also noticed examples of variation in pronunciation across different speakers in the collection and stated that a useful lesson would be to point out these differences to students to raise their awareness of word stress, intonation, and connected speech in spoken Hawaiian. In their view, using the tool for pronunciation practice could help reduce students' stress when speaking in front of others. A second teacher stated that using the KWIC feature to generate numerous examples of a specific word or structure can allow teachers to design more focused activities for pronunciation and speaking practice.

7.2 User Training

In the second phase of the project, information on learner needs and user feedback collected in the first phase will be incorporated into teacher training workshops focusing on how to use the corpus toolkit for designing teaching materials and how to incorporate the tool in the classroom for data-driven learning. The goals of the workshops are to help build teacher corpus literacy, introduce participants to corpus-based language pedagogy (Ma et al., 2024), and guide participants in designing teaching materials using the tool. Training sessions on using the toolkit for data-driven learning will also be conducted with individual Hawaiian learners, with the goal of enabling participants to utilize the platform for autonomous learning, with a focus on listening and pronunciation practice. User feedback from teachers and learners will also be gathered in this phase to inform further improvement of the platform.

8 Conclusion

We have presented an online corpus toolkit for data-driven language learning of Hawaiian. Designed for teachers and students, but with many potential applications, this toolkit supports several corpus linguistics analyses including concordances, frequency, collocations, and complex n-gram and regex searches, which can greatly facilitate the language learning process. One important feature of our corpus is the inclusion of audio, which enables students to learn pronunciation from native speakers of Hawaiian. The technical implementation was done with both simplicity and efficiency in mind, leveraging PostgreSQL and an inverted index data structure for fast querying. The toolkit is currently online for internal usage, and there are plans to make it publicly available once it is more mature. We have conducted interviews with experienced Hawaiian teachers about their needs and how they can effectively use the toolkit in their classrooms. We are in the process of organizing the second phase of community-driven corpus development, where we will host training sessions for Hawaiian teachers, students, and parents to use this toolkit.

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 au-

thors and do not necessarily reflect the views of the NSF. The authors would also like to thank Ha‘alilio Solomon and Laiana Wong for participating in this study by providing valuable feedback.

Ethical Considerations and Limitations

The corpus collected in this paper comes from Ulukau, the Hawaiian Electronic Library. The data is available for educational purposes, and explicitly forbids commercial usage. In terms of limitations, our toolkit currently has limited testing by Hawaiian learners, as we have been developing the toolkit in consultation with Hawaiian teachers. Nevertheless, existing literature has shown that corpus tools promote data driven learning, and we expect to continue the development of our toolkit with more feedback from teachers and students in the future.

References

Laurence Anthony. 2005. Antconc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005.*, pages 729–737. IEEE.

Christopher M Baker. 2012. *A-class genitive subject effect: A pragmatic and discourse grammar approach to a-and o-class genitive subject selection in Hawaiian*. Ph.D. thesis, University of Hawaii at Manoa.

Alex Boulton and Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language learning*, 67(2):348–393.

Vaclav Brezina and Dana Gablasova. 2015. Is there a core general vocabulary? introducing the new general service list. *Applied Linguistics*, 36(1):1–22.

Catherine Elizabeth Lee Brockway. 2021. *Building high-frequency word lists for the semantic domain of ‘ĀINA (‘land’) using a raw corpus of spoken ‘ōlelo Hawai‘i*. Ph.D. thesis, University of Hawai’i at Manoa.

Maggie Charles. 2022. Corpora and autonomous language learning. In *The Routledge handbook of corpora and English language teaching and learning*, pages 406–419. Routledge.

Matthew Coole, Paul Rayson, and John Mariani. 2020. LexiDB: Patterns & methods for corpus linguistic database management. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3128–3135, Marseille, France. European Language Resources Association.

William J Crawford. 2022. Corpora and speaking skills. In *The Routledge handbook of corpora and English language teaching and learning*, pages 89–101. Routledge.

Niall Curry and Geraldine Mark. 2024. Using corpus linguistics in materials development and teacher education. *Second Language Teacher Education*, 2(2):187–208.

Niall Curry and Tony McEnery. 2025. Corpus linguistics for language teaching and learning: A research agenda. *Language teaching*, pages 1–20.

Ewa Czaykowska-Higgins. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within canadian indigenous communities. *Language Documentation & Conservation*, 3(1):15–50.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

Benjamin Frey. 2018. “Data is nice:” Theoretical and Pedagogical Implications of an Eastern Cherokee Corpus. *Language Documentation I& Conservation Special Publication*, 20:38–53.

Dee Gardner and Mark Davies. 2014. A new academic vocabulary list. *Applied linguistics*, 35(3):305–327.

Emily‘Ioli‘i Hawkins. 2003. Distribution and function of hawaiian ana. In *Rongorongo Studies: A Forum for Polynesian Philology*, volume 13, pages 3–19.

Kelsea Kanohokuahiwi Hosoda. 2019. *Hawaiian morphemes: Identification, usage, and application in information retrieval*. Ph.D. thesis, University of Hawai’i at Manoa.

Tim Johns. 1991. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *English Language Research Journal*, 4:27–45.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine. *Lexicography*, 1(1):7–36.

Larry Kimura. 2025. Kani‘āina: Ka Leo Hawai‘i Collection. `https://ulukau.org/kaniaina`. [Accessed 21-10-2025].

Dawn Knight and Svenja Adolphs. 2021. Multimodal corpora. In *A practical handbook of corpus linguistics*, pages 353–371. Springer.

Pia Lane, Kristin Hagen, Anders Nøklestad, and Joel Priestley. 2022. Creating a corpus for kven, a minority language in norway. *Nordlyd*, 46(1):159–170.

Qing Ma, Rui Yuan, Lok Ming Eric Cheung, and Jing Yang. 2024. Teacher paths for developing corpus-based language pedagogy: A case study. *Computer Assisted Language Learning*, 37(3):461–492.

Tony McEnery and Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Anne O'Keeffe and Michael McCarthy. 2022. *The Routledge handbook of corpus linguistics*, volume 10. Routledge London.

Anne O'Keeffe, Michael McCarthy, and Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Harold Edward Palmer. 1933. Second interim report on english collocations, submitted to the tenth annual conference of english teachers, under the auspices of the institute for research in english teaching.

Pascual Pérez-Paredes. 2022. A systematic review of the uses and spread of corpora and data-driven learning in call research during 2011–2015. *Computer Assisted Language Learning*, 35(1-2):36–61.

Keren Rice. 2018. Collaborative research: Visions and realities. In *Insights from practices in community-based research: From theory to practice around the globe*, pages 13–37. Mouton de Gruyter Berlin, Boston.

Sally Rice and Dorothy Thunder. 2017. Community-based corpus-building: Three case studies. In *ICLDC-5*, University of Hawai'i at Mānoa, Honolulu, HI.

Ute Römer. 2011. Corpus research applications in second language teaching. *Annual review of applied linguistics*, 31:205–225.

John M. Sinclair. 2004. *How to use corpora in language teaching*. John Benjamins Publishing Company.

Daniela Teodorescu, Josie Matalski, Delaney Lothian, Denilson Barbosa, and Carrie Demmans Epp. 2022. Cree corpus: A collection of nêhiyawêwin resources. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6354–6364, Dublin, Ireland. Association for Computational Linguistics.

Kepano Trussel. 2022. CHD - Combined Hawaiian Dictionary. `https://www.trussel2.com/HAW/`. [Accessed 21-10-2025].

Amy Tsui. 2004. What teachers have always wanted to know—and how corpora can help. inj. sinclair (ed.), how to use corpora in language teaching (pp. 39–61).

Nina Vyatkina. 2020. Corpora as open educational resources for language teaching. *Foreign Language Annals*, 53(2):359–370.

Hongzhi Xu, Helen Kaiyun Chen, Chu-Ren Huang, Qin Lu, Dingxu Shi, and Tin-Shing Chiu. 2012. A grammar-informed corpus-based sentence database for linguistic and computational studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3140–3144, Istanbul, Turkey. European Language Resources Association (ELRA).