Bridging Digital Tools for Linguistic Documentation and Revitalization

Christopher Haberland* **Carly Crowther*** **Jingnong Qu** **Anuk Centellas**
University of Washington
{haberc, carlyc88, jingnong, anukz}@uw.edu
*Joint first authors.

Abstract

Digital tools serving language revitalization tend to fall into two categories: 1) linguist-oriented documentation tools that prioritize annotation, morphological analysis, and archival preservation, and 2) community-facing applications that emphasize accessibility and language learning. Few systems integrate the former with the latter, and practical barriers — including the cost of computational expertise, single-user workflows, and limited data governance — further constrain their utility. These disconnects incur additional development and communication costs for revitalization teams consisting of linguists and community members. We introduce `langlit`, a collaborative web-based platform that attempts to tailor documentation workflows for the language revitalization context within a single system. The platform integrates a finite-state morphological analyzer with a three-tier human-in-the-loop annotation workflow, searchable corpus interfaces with multiple query modalities, interactive word construction guided by the morphological grammar, corpus-linked hypothesis tracking with provenance, and a grammar-derived editable dictionary. All components share a single underlying FST grammar, and the system supports configurable access controls, collaborative editing, and optional LLM integration with transparent data handling. Designed for redeployment across languages through a modular architecture, `langlit` is published as an open-source repository on GitHub. We situate our system within the existing landscape of revitalization tools through a comparative analysis and discuss how integrated, community-informed design can better serve the specific goals of language revitalization.

1 Introduction

Digital tools for language revitalization and language documentation serve different goals. Documentation tools prioritize annotation, archival standards, and linguistic analysis, while revitalization efforts focus on intergenerational transmission and everyday language use (Flavelle and Lachler, 2023; Gessler, 2022). Software designed principally for documentation may lack features desired by revitalization teams whose work involves not only collating linguistics knowledge, but also transmitting it in ways consistent with community values (Le Ferrand et al., 2022; Gessler, 2022). As a result, linguistic knowledge produced during documentation often remains inaccessible to teachers and learners (Neubig et al., 2020).

Practical barriers compound this disconnect. Documentation tools are largely unfamiliar outside academic linguistics (Skilton et al., 2025), and building accessible, community-facing applications requires developer time that most revitalization projects cannot sustain (Wagner, 2017). Few reusable templates exist that communities could adapt independently.

In response, we introduce `langlit`, an open-source web-based system designed to bridge documentation and revitalization workflows. The system makes annotations and corpus data accessible to teachers and learners through collaborative editing, corpus search, and configurable data governance. It is designed so that the products of linguistic documentation are immediately and transparently available to language community stakeholders as documentation work progresses.

Beyond its organizational and collaboration features, the system is designed to hasten linguistic knowledge discovery. A graphical interface allows for generation of morphological interpretations of a text, and optional tooling integrating large language models is provided to assist during the processes of annotation and linguistic discovery. The design emphasizes language generality and modularity so that it can be adapted for other efforts. We invite open contributions to our work, which is

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 22–32

published as an open repository on GitHub. [1]

Section 2 reviews the landscape of digital tools for language documentation and revitalization. We examine their features, limitations, and how they emphasize certain roles over others. Section 3 describes the system architecture and its core components. Section 5 discusses limitations and future work.

2 Background

Structured corpora, morphological parsing, and interlinear glossing are rarely incorporated into tools intended for teachers and learners, despite broad recognition that revitalization efforts benefit from example-rich, searchable materials. This section reviews the features and limitations of existing tools to motivate the design of systems that bridge documentation and community use.

2.1 Features of digital tooling for language revitalization

Digital tools expand both the reach and flexibility of language revitalization efforts. Online platforms, mobile applications, and multimedia resources enable geographically dispersed learners to engage with a language outside of traditional classroom settings, which is particularly valuable for diaspora communities lacking in-person access to fluent speakers (Chew et al., 2023; Mauger, 2025; Meighan, 2024). Against the backdrop of reduced intergenerational transmission, language technologies can support classroom teaching by providing accessible, searchable corpora and recordings of fluent speakers (Mainzinger, 2024). Multimedia materials further support culturally grounded learning by pairing linguistic data with narratives and oral histories (Meighan, 2021).

A central benefit of language technology is its ability to distribute authentic language materials across time and space, supporting learning even where fluent speakers are few or geographically dispersed (Richards et al., 2025; Tennell and Chew, 2024). These considerations motivate the design of systems that combine analytical rigor with accessibility, collaboration, and pedagogical support.

2.2 Linguist documentation tools

The most prominent documentation tools are technical environments developed by and for linguists. FieldWorks Language Explorer (FLEx) integrates a lexicon, text corpus, and morphological parser within a unified workflow, making it widely used in language documentation , and no other tool combines these three components in an integrated fashion (Skilton et al., 2025) (see Table 1). Another widely used tool, ELAN, supports multi-tier annotations of audio and video (Wittenburg et al., 2006), but its workflow assumes significant linguistic expertise and privileges researchers' needs over those of teachers or community members (O'Neil et al., 2024). ELAN's interlinearization features remain under development and do not yet provide the integrated lexicon–parser workflow valued in FLEx (Skilton et al., 2025).

Despite their important role, these tools have notable limitations. FLEx is difficult to install and operate without specialized training, is optimized for single-user workflows with limited version control, and lacks transparent version histories, complicating accountability for changes to language data (Skilton et al., 2025). Its management by SIL International and the requirement for data to be uploaded to an outside server raises data privacy and sovereignty concerns (Skilton et al., 2025). Both tools require powerful local computing and a desktop workflow.

The technological landscape has evolved significantly since these tools were first developed. Smartphones, web applications, and cloud-based collaboration are now widespread, and advances in natural language processing have expanded what is possible for both linguists and learners. However, few of these advances have been used to upgrade core documentation software.

Integrating NLP models into tools like FLEx and ELAN remains technically cumbersome, to the point of discouraging adoption even among experienced documentary linguists (Gessler, 2022). This suggests a need for systems that retain the analytical depth of existing documentation tools while supporting modern, collaborative workflows for the downstream users of language documentation outputs.

2.3 Community-facing tools

Community-facing tools such as mobile applications, web dictionaries, gamified learning tools, and multimedia storytelling platforms are intended for direct use by language communities (Ajani et al., 2024; Bettinson and Bird, 2021; Galla, 2016; Tennell and Chew, 2024). In communities with

[1] https://github.com/haberchr/langlit

few fluent speakers, digital platforms distribute recordings of Elders, extending the reach of limited linguistic resources (Meighan, 2021, 2024). For example, Littell et al. (2017) introduce a reusable framework for web dictionaries across languages and Kazantseva et al. (2018) describe a verb conjugation tool for Kanyen'keha designed to help learners navigate complex morphology. These tools are often effective for engagement and beginner learning but few consistently integrate morphological search functionality over annotated corpora, limiting utility for teachers and advanced learners who need contextualized examples of specific grammatical phenomena (Neubig et al., 2020; Taylor-Adams, 2019).

Beyond functional limitations, concerns about data sovereignty are central to the design of community-facing language technologies (Chew et al., 2023; Kukutai and Taylor, 2017; Schwab-Cartas, 2018; Tennell and Chew, 2024). These commitments are difficult to uphold when projects rely on external infrastructure such as third-party platforms or large language models. Few reusable, community-adaptable application frameworks exist, so each new project typically requires full custom development. This is a structural barrier because communities with the greatest need are often the least positioned to commission or maintain such tools (Chew, 2021; Wagner, 2017).

Despite their complementary strengths, documentation and community-facing tools have largely developed in parallel, and as Gessler (2022) notes, existing software has not kept pace with the collaborative, accessible workflows that revitalization teams need. Bridging this gap requires platforms with usable interfaces, collaborative capabilities, and community governance over language data.

2.4 Comparison of select work

Table 1 compares the documentation, pedagogical, and infrastructure features of a sample of work identified in our review of digital tools for language documentation and revitalization. Our review illustrates the general patchwork of features offered by tools purposed for revitalization that motivated the design of `langlit`.

Neither FLEx nor ELAN, the most widely used documentation applications, provide a pedagogical interface for language community members. FLEx offers concordance generation and regex filtering within its Texts & Words module, but this does not permit search functionality accessible to non-specialists (Skilton et al., 2025). SIL has recently released FieldWorks Lite (beta)[2], a cross-platform companion application that introduces real-time collaborative editing for lexicon data, though it does not support morphological parsing, text annotation, or the broader documentation workflows provided by FLEx. ELAN supports multi-tier regex search over annotated media, but a documented limitation in its Multiple Layer Search prevents reliable morphosyntactic queries across utterance boundaries, restricting its utility as a corpus search tool (Wilbur, 2019). Neither tool offers integrated word construction, collaborative hypothesis documentation, or configurable data governance.

Gessler (2022) presents Glam, a work-in-progress system that most directly shares `langlit`'s goal of bridging NLP models and documentary linguistics through shared software infrastructure. Glam targets documentary linguists and NLP integration, providing UIs for annotation and model interaction, but does not describe a pedagogical interface for teachers or learners. Plaid[3] constitutes a linguistic annotation backend that is designed to solve data management and collaboration for linguistic documentation applications, but is not deeply coupled with analytical or pedagogical tooling.

Among the community-facing tools in our comparison, Littell et al. (2017) and Debenport et al. (2023) both offer dictionary interfaces with lexical search, but neither supports morphological corpus search over annotated texts. Mukurtu, described by Debenport et al. (2023), is the most developed system for data governance, providing granular, protocol-based access control designed specifically for cultural sovereignty needs; however, Mukurtu supports metadata and media annotation within its archival framework rather than linguistic corpus annotation such as morphological glossing or interlinear glossed text. Richards et al. (2025) describe a mobile application with corpus search and multimodal features, but without collaborative editing, open-source availability, or morphological analysis. Kazantseva et al. (2018) stands out as the only system besides `langlit` that provides both a morphological analyzer and an

[2] https://software.sil.org/fieldworks/download/fieldworks-lite/

[3] https://larc-iu.github.io/plaid/manual.html

	Documentation and pedagogical capabilities									Infrastructure & governance				
Tool	Morph. Analyzer	Dict.	Corpus Search	Pedagog. UI	Corpus Annot.	Multi-modal	Phrase Builder	Hypoth. Doc.	LM Integ.	Collab. Editing	User Mgmt	Open Source	Data Gov.	Platform
FLEx	✓	✓	*limited	—	✓	—	—	—	—	—	*limited	✓	—	Desktop
ELAN	—	—	*limited	—	✓	✓	—	—	—	—	*single	✓	*?	Desktop
Littell et al. (2017)	—	✓	*lex	✓	—	—	—	—	—	—	—	✓	*?	Web/Mobile
Kazantseva et al. (2018)	✓	—	—	✓	—	—	✓	—	—	—	—	—	—	Web/Mobile
Gessler (2022)	—	—	—	—	✓	—	—	—	✓	✓	✓	✓	*✓[a]	Web
Debenport et al. (2023)	—	✓	*lex	✓	*limited	✓	—	—	—	✓	✓	✓	✓	Web
Pugh and Tyers (2023)	✓	—	—	—	—	—	—	—	—	—	—	✓	—	CLI
Mainzinger (2024)	—	✓	—	✓	—	*?	—	—	—	—	—	—	—	*mixed
Cox et al. (2025)	✓	✓	—	✓	—	✓	—	—	—	—	—	✓	*?	Web
Richards et al. (2025)	—	—	✓	✓	—	✓	—	—	—	—	—	—	*token	Mobile
Hammerly et al. (2026)	✓	✓	—	*ext	—	—	*ext	—	—	—	—	✓	—	CLI
`langlit`	✓	✓	✓	✓	✓	—	✓	✓	✓	✓	✓	✓	✓	Web

Table 1: Comparison of selected language revitalization tools. ✓ present; — absent; * partial or unclear (asterisk with label). *lex*: search limited to lexical lookup. *limited*: feature present but with significant constraints. *ext*: downstream application described but not integrated into a single UI. *single*: single-user only. *token*: token/password-based access. *?*: unclear from available documentation. *mixed*: no single integrated platform. LM Integ. refers to in-app integration with language models (e.g., LLMs, neural taggers) for annotation assistance or other tasks, not rule-based morphological parsing. *(a) Derivative work `Plaid` adeptly addresses infrastructure and governance issues.

integrated word/phrase construction interface for learners; however, it lacks corpus search, annotation, and collaborative infrastructure.

Several recent projects build directly on FST-based morphology. Pugh and Tyers (2023) describe a finite-state analyzer for Highland Puebla Nahuatl, but without an accompanying application layer, and Mainzinger (2024) presents a technology roadmap for Mvskoke drawing on multiple existing systems rather than a single integrated platform. Cox et al. (2025) describe a long-term community partnership to build an FST-backed intelligent dictionary for Tsuut'ina, with verb paradigms reviewed item-by-item by a community language committee. Hammerly et al. (2026)'s Ojibwe-Morph introduces an FST framework whose downstream applications include a verb conjugation tool for education, a spell-checker, and intelligent dictionary search. In each case, pedagogical and word-construction capabilities are built on FST output, but do not provide corpus annotation, corpus search, or collaborative editing capabilities.

Among the tools surveyed, no system combines a morphological analyzer, corpus annotation, corpus search, a pedagogical interface, word construction, and hypothesis documentation within a single collaborative, open-source, web-based platform with configurable data governance. `langlit` is designed to address this gap.

3 System Description

`langlit` is a Python application built with the `streamlit`[4] framework used for displaying data-centric applications as web interfaces. It is easily shareable online via Streamlit Community Cloud with minimal setup. `langlit` integrates an FST morphosyntactic grammar with corpus annotation, search, word construction, linguistic hypothesis tracking, and a dictionary. The user-defined FST grammar and phonological file backend (`.lexc`/`.xfscript`) are read by the Helsinki Finite State Transducer Python package `hfst` (Lindén et al., 2011) and and form the backbone of the morphological interpretation engine.

The system can be customized for a target language by creating a "language pack." This consists of a configuration module defining language-specific parameters that are declared in a single `config.py` file imported by all pages, along with other feature and language specific files to support the morphological analysis backend. This architecture supports cross-language applicability. Communities can enable or disable components as needed for the intended audience (linguists or community teachers and learners) or desired functionality.

In the following subsections, we describe the system's core components: human-in-the-loop annotation workflows (§3.1), multi-modal corpus

[4] `https://streamlit.io`

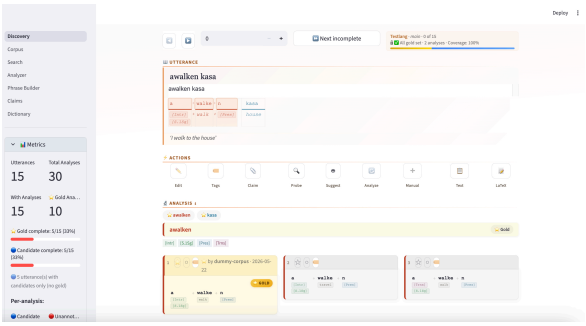

Figure 1: The `langlit` Discovery page.

search (§3.2), IGT export and pedagogical material generation (§3.3), corpus-linked hypothesis tracking (§3.4), a grammar-derived editable dictionary (§3.5), and collaborative editing with configurable access control (§3.6).

3.1 Human-in-the-Loop Annotation Workflows

The Discovery page (Figure 1) serves as the primary interface for corpus exploration and annotation management. It provides an overview of corpus-level metadata and annotation progress, including utterance analyses and gold annotation counts, allowing users to assess the state of the corpus at a glance.

For each utterance, Interlinear Glossed Text (IGT) is rendered inline as an HTML table aligned to the tokenized surface form, with spans lacking a gold annotation (*gaps*) highlighted in red to make annotation completeness visually salient. The table displays the normalized surface form, morpheme boundaries, and gloss. Utterances can be navigated sequentially or by jumping directly to incomplete entries, streamlining the annotation workflow. Critically, the underlying data model also supports multilingual comparison of utterances, addressing a documented limitation of FLEx (Skilton et al., 2025).

The sidebar computes and displays annotation progress metrics, including per-analysis counts of gold, candidate, and unannotated spans, and the rate of agreement between gold and candidate labels across all analyses with both labels set. Disagreements are flagged for reconciliation. A "next incomplete" navigation control moves the reviewer directly to the next utterance lacking a complete gold annotation, simplifying navigation through a large corpus.

The Discovery page is primarily designed for linguists, researchers, and annotators who need to systematically understand the morphosyntactic properties of a corpus and build a gold-annotated dataset that can be formatted as IGT. It also serves as a useful exploration tool for teachers and advanced learners seeking a deep understanding of the language's structure. Once labeled, the data becomes a valuable resource for teachers who may use the Search page to identify relevant linguistic examples for lesson creation.

3.1.1 FST-assisted analysis

An FST analyzer supports corpus review on the Discovery page. Finite-state transducers are well-suited for morphological analysis in revitalization contexts because they do not require large training corpora, which are often unavailable for low-resource languages (Kazantseva et al., 2018). Unlike neural approaches requiring expensive training runs and expansive amounts of digitized data resources, FSTs are rule-based, deterministic, and can be constructed directly from morphosyntactic knowledge. This allows for early-stage and continuous development as the documentation team learns more about the language, even in the context of resource scarcity. Many Indigenous languages have complex inflectional systems yielding extremely large numbers of surface forms (Mithun, 2001), making the concatenative approach (Hammerly et al., 2026) that `langlit` adopts particularly appropriate. FST-based parses discretize morphological transitions, keeping analysis transparent and correctable—especially important where analyses must remain interpretable by less technically inclined team members (Pugh and Tyers, 2023).

Beyond morphological analysis, the software used to create FST-based systems is highly extensible to a range of downstream modular applications, including spell checkers, grammar checkers (Pirinen et al., 2023), verb conjugators (Kazantseva et al., 2018), interactive transcription systems (Lane and Bird, 2022) and morphologically aware dictionary search (Hammerly et al., 2026; Cox et al., 2025). The app realizes several of these extensions directly: the Phrase Builder page uses the same `.lexc` grammar to guide interactive word construction (§3.3), and the Dictionary page derives its lexeme and morpheme inventory from the same source (§3.5).

3.1.2 Three-tier annotation workflow

Annotation enrichment is expensive, and revitalization programs rarely have the resources required to annotate a corpus from scratch (Nicolai et al., 2020; Neubig et al., 2020). The Discovery page implements a three-tier human-in-the-loop design. The first tier is human-labeled annotation, supporting tabula rasa span annotation for morphological or other properties according to user-defined conventions.

The second and third tiers are motivated by recent work that shows the potential for NLP tooling to improve the speed at which high quality annotations can be produced (Ginn et al., 2024; Liang et al., 2026). These tiers are optional to allow teams to annotate according to their preference and utility. The second tier allows morphological annotation using finite state transducers via `hfst` (Lindén et al., 2011), a package that references `.lexc` and `.xfscript` files to deterministically infer morphological annotations for spans of text.

The third tier, useful in parallel data contexts, prompts a server-based or locally hosted large language model (LLM) to distinguish from among morphological interpretations generated by the FST step to choose the most explanatory interpretation given its parallel translation of the referenced data in a high-resource language.

It is important to highlight that many groups working towards revitalization may prefer to not send any target language data to AI servers to maintain maximal data ownership and sovereignty; to this end, our example workflow does not provide any surface-form target language data to the AI server in this step - only the morphological glosses are passed. The external API call is opt-in, gated behind a UI and password toggle, and can be disabled entirely by communities that prefer fully manual review or have concerns about data exposure through third-party services, aligning with data privacy principles addressed in O'Neil et al. (2024). Indeed, data sovereignty concerns may also be allayed by selecting local LLM models for inference in this step, if desired.

A human reviewer may assign a *gold* label that may agree with or override *candidate* interpretations suggested by the FST and selected by then LLM. The UI ensures that annotations labeled by humans and machine systems are always distinguished and apparent. Declined analyses are tracked separately from unannotated ones.

3.2 Multi-Modal Corpus Search

Documentary corpora contain naturalistic, morphologically annotated examples drawn from actual usage, yet these remain practically out of reach for teachers without specialist training (Neubig et al., 2020; Taylor-Adams, 2019).

Low-resource language corpora "typically lack not only graphical search interfaces, but also the rich annotations (such as morphological and syntactic parses) that are conventionally required to support the function of a search interface" (Neubig et al., 2020). Compounding this, search interfaces that do exist typically require users to query in technical terms that are beyond the expertise of most teachers and learners (Taylor-Adams, 2019).

The Search page allows access to corpus entries containing a specified word or phrase, enabling users to efficiently locate and examine instances of target forms in context. Search results display the full utterance alongside its translation. Entries can be bookmarked and downloaded in CSV, PDF, or LaTeX format, making it straightforward to compile collections of examples for further analysis or instructional use.

Entries can be searched by translations, morphological glosses, tags, or a combination. Beyond the simple lexical lookup offered by other tools (see Table 1), the Search page supports regex and neural embedding queries over both monolingual and bilingual corpora, enabling comparative analysis and evidence gathering for linguistic research and teaching.

For linguists and researchers, the Search page provides a fast way to identify and examine instances of specific morphological phenomena across the corpus. It simplifies the process of finding examples with specific grammatical features or vocabulary, supporting the preparation of lesson materials such as handouts and study guides, which is useful for teachers. The Search page also offers an accessible entry point for learners to explore how particular words or concepts appear in the language.

All modes return paginated example cards displaying the surface form, gloss, and IGT of the gold analysis. Results of a search can be individually bookmarked within a session and staged for linking to a **hypothesis** being evaluated by a linguistic research team(§3.4), or immediately exported in a CSV for offline use. This design directly re-

sponds to the Teacher-in-the-Loop model of Neubig et al. (2020), which called for corpus retrieval that does not require users to express queries in technical terms and can function over partially annotated data.

3.3 IGT Export and Pedagogical Material Generation

Language documentation has historically prioritized researcher access over community-usable resource creation (Gessler, 2022). The app addresses this gap with two pages that convert annotation and grammar resources directly into exportable pedagogical materials.

3.3.1 Analyzer

The Analyzer page provides an interactive interface for morphological analysis of utterances. It accepts free-text input in the target language and runs it through the same preprocessing and FST pipeline. Users can examine the full set of interpretations for each span and designate one as the gold standard analysis, building an IGT entry for a word or phrase in real time.

Users can bookmark analyzed utterances, which are added to an export queue that can be downloaded, enabling ready-to-use IGT displays for external documents. The resulting three-line interlinear (surface / morphological / gloss) is exported as a standalone LaTeX document using a tabular layout compatible with standard linguistic IGT conventions, or as a PDF via a `pdflatex` subprocess call. An optional LLM call proposes a free translation from the gold gloss line.

The Analyzer is primarily aimed at linguists and researchers who need to inspect and validate FST output for specific forms. It is also a valuable tool for learners seeking a deeper understanding of the morphological structure of individual words, and for teachers looking to create IGT-based exercises or materials for translation and morphological analysis practice.

3.3.2 Phrase Builder

The Phrase Builder page provides a morphologically guided phrase construction interface that traverses the `.lexc` FST while exposing the FST's morphological generation functionality in a user-understandable way. Users begin by selecting a part of speech and searching for a specific root meaning, then are guided through answers to a series of grammatical questions that iteratively build a morphologically complex word from a root meaning. The interface reformulates each morphological choice as a grammatical question (e.g., a dropdown menu for person-number agreement), using question templates and tag-to-question mappings defined in a configuration file and comments in the `.lexc` file. `lexlit` currently supports adding phonological and infix rules via Python or `.xfscript`.

The Phrase Builder serves a range of users. For linguists and researchers, it provides an environment for exploring the productive morphology of the language. For learners, the guided step-by-step interface offers an accessible way to develop intuitions about the morphological structure of the language. For teachers, the page supports the construction of paradigms and morphologically varied word forms that can be exported and used as lesson materials.

3.4 Corpus-Linked Hypothesis Tracking

Neither the tools reviewed by Neubig et al. (2020) and O'Neil et al. (2024) nor any tool in Table 1 provide a mechanism to link specific corpus examples to in-progress linguistic hypotheses. The Claims page addresses this gap by implementing a structured system for documenting and tracking linguistic hypotheses or discoveries about the language. From both the Discovery and Search pages, any corpus utterance can be linked to a claim with an evidence relationship (*supports*, *contradicts*, or *neutral*), along with a mandatory note explaining the relevance of the example and a username and timestamp for provenance. A claim detail view displays the full evidence set and tracks supporting and contradicting evidence counts separately. Every update to a claim generates a timestamped revision record, providing an audit trail of how analysis evolved. Claims can be exported as structured LaTeX or PDF documents displaying the claim and associated evidence to support pedagogical communication.

Because evidence links are attributed to named users and claim status is visible to all collaborators, the system supports the transparent, reciprocal research process called for by O'Neil et al. (2024), which is particularly valuable for underdocumented languages where hypotheses may need to be revised as continued analysis of a corpus yields new insights.

The Claims page is a useful reference for teachers who want to incorporate verified linguistic

knowledge into lesson planning and for independent learners seeking a deeper understanding of the language's structure. Future work will 1) document agentic AI tools for automating linguistic discovery using a custom `langlit` harness, and 2) improved grammar reference distribution tools for members of the language community.

3.5 Grammar-Derived Editable Dictionary

A recurring problem in documentation is that dictionary materials exist as static files disconnected from the morphological resources used for analysis (Neubig et al., 2020; Cox et al., 2025). The Dictionary page addresses this by deriving its content directly from the same `.lexc` file used by the FST and Phrase Builder. Because the dictionary is generated from the same resource that powers analysis, there is no structural divergence between dictionary coverage and FST coverage: adding a morpheme to the analyzer makes it immediately visible in the dictionary. This tight coupling eliminates the need to keep a separate lexical database synchronized with a changing analyzer.

The Dictionary page displays each entry alongside a translation. Users can search for specific entries, filter by lexicon, and group results by lexicon or gloss, making it straightforward to navigate and organize the vocabulary of the language. Entries can be edited directly, allowing the lexical database to be updated as new information is discovered. Users may edit the categories from the `.lexc` file that are displayed via the language pack `.yaml` configuration file.

The Dictionary provides linguists and researchers an organized, searchable view of the lexicon with the ability to make edits. For teachers and learners, it functions as a practical vocabulary reference, enabling quick lookup of specific words and their translations for use in lessons or independent study.

3.6 Collaborative, Transparent Editing with Configurable Access Control

O'Neil et al. (2024) identify collaborative editing, user management, edit history, and data transparency as essential features of any cross-culturally applicable documentation tool, specifically noting that FLEx's single-user workflow and opaque version history as significant limitations (Skilton et al., 2025).

3.6.1 Web deployment

The application is deployable as a web application via Streamlit Community Cloud or other providers. For development purposes, computational linguists and developers may easily deploy the application locally or on custom servers.

3.6.2 Access control

User identity is established through OpenID Connect (OIDC) authentication on Streamlit Community Cloud and in local deployments after server configuration. Access is gated by Streamlit's authentication function (`st.login()`), which executes an OICD flow with Google as the identity provider, restricting app access to admin-defined accounts. Editing and LLM-usage is gated behind an admin-defined password, so non-privileged users have read-only access by default. The app currently supports SQLite as a backend, which is backed up to a Google Workspace account established by the admin. The local database allows for concurrent read access and serialized writes for users of the app. Future work could support cloud-native storage solutions.

3.6.3 LLM Integration and data transparency

External API calls are entirely opt-in: the LLM toggle's function is transparently presented to the user, and users may opt to avoid passing language data through third-party services by disabling any LLM inference without affecting other tool functionality.

4 Conclusion

Digital tools for language documentation and revitalization have historically directly served either linguists or community members, but rarely both. To address this gap, we present `langlit`, an open-source, web-based platform that combines a finite-state morphological analyzer, three-tier human-in-the-loop annotation workflow, multi-modal corpus search, interactive word construction, corpus-linked hypothesis tracking, and a grammar-derived dictionary in a single system. Because all components draw from one shared FST grammar, adding a morpheme to the grammar immediately propagates across the analysis, search, word construction, and dictionary pages. The platform is designed for redeployment across languages through a modular architecture, and its configurable access

controls reflect the data sovereignty priorities common in Indigenous language contexts. We hope that `langlit` can lower barriers for linguists and language community members to create, access, and use language documentation to benefit revitalization work.

5 Limitations

A major limitation of `langlit` is that its customization and integration in a language revitalization context may require involvement of a developer or computational linguist. We acknowledge that this assumption falls short of many real-world contexts and remains a problem to be addressed by future work.

We have not yet conducted user acceptance testing among various stakeholder roles identified in this work. Sustained use from users of varied perspectives is critical for corroborating many of the design assumptions introduced in this work.

We acknowledge that the software presented does not yet incorporate all desiderata for language documentation, analysis, and teaching. For example, our application does not currently support the inclusion of multimodal annotations. It is our hope that the application may serve as a starting point to allow others to suit the software to their needs.

References

Yusuf Ayodeji Ajani, Bolaji David Oladokun, Shuaib Agboola Olarongbe, Margaret Nkechi Amaechi, Nafisa Rabiu, and Musediq Tunji Bashorun. 2024. Revitalizing Indigenous Knowledge Systems via Digital Media Technologies for Sustainability of Indigenous Languages. *Preservation, Digital Technology & Culture*, 53(1):35–44.

Mat Bettinson and Steven Bird. 2021. Collaborative fieldwork with custom mobile apps. *Language Documentation & Conservation*, 15:411–432.

Kari A. B. Chew. 2021. #KeepOurLanguagesStrong: Indigenous Language Revitalization on Social Media during the Early COVID-19 Pandemic. *Language Documentation and Conservation*, 15:239–266.

Kari A.B. Chew, Sara Child, Jackie Dormer, Alexa Little, Olivia Sammons, and Heather Souter. 2023. Creating Online Indigenous Language Courses as Decolonizing Praxis. *The Canadian Modern Language Review*, 79(3):181–203.

Christopher Cox, Bruce Starlight, Janelle Crane-Starlight, Hanna Big Crow, and Antti Arppe. 2025. Creating an intelligent dictionary of tsuut'ina one verb at a time. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 110–119, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Erin Debenport, Mishuana Goeman, Maria Montenegro, and Michael Wynne. 2023. How a Dictionary Became an Archive: Community Language Reclamation Using the Mukurtu Content Management System. *Dictionaries: Journal of the Dictionary Society of North America*, 44(2):29–55.

Darren Flavelle and Jordan Lachler. 2023. Strengthening Relationships Between Indigenous Communities, Documentary Linguists, and Computational Linguists in the Era of NLP-Assisted Language Revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Candace K. Galla. 2016. Indigenous language revitalization, promotion, and education: function of digital technology. *Computer Assisted Language Learning*, 29(7):1137–1151.

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 Ginn, Mans Hulden, and Alexis Palmer. 2024. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Hammerly, Nora Livesay, Antti Arppe, Anna Stacey, and Miikka Silfverberg. 2026. OjibweMorph: An Approachable Finite-State Transducer for Ojibwe (and Beyond). *Language Resources and Evaluation*, 60:27.

Anna Kazantseva, Owennatekha Brian Maracle, Ronkwe'tiyóhstha Josiah Maracle, and Aidan Pine. 2018. Kawennón:nis: the wordmaker for Kanyen'kéha. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 53–64, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tahu Kukutai and John Taylor. 2017. *Indigenous Data Sovereignty: Toward an agenda*. ANU Press.

William Lane and Steven Bird. 2022. A finite state aproach to interactive transcription. In *Proceedings of the First Workshop on NLP applications to field linguistics*, pages 1–10, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Fashioning local designs from generic speech

technologies in an Australian aboriginal community. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4274–4285, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Siyu Liang, Talant Mawkanuli, and Gina-Anne Levow. 2026. Hybrid Neural-LLM Pipeline for Morphological Glossing in Endangered Language Documentation: A Case Study of Jungar Tuvan. In *Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics*, pages 16–30, Rabat, Morocco. Association for Computational Linguistics.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A. Pirinen, and Miikka Silfverberg. 2011. HFST—Framework for Compiling and Applying Morphologies. In *Systems and Frameworks for Computational Morphology*, pages 67–85, Berlin, Heidelberg. Springer.

Patrick Littell, Aidan Pine, and Henry Davis. 2017. Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 141–150, Honolulu. Association for Computational Linguistics.

Julia Mainzinger. 2024. Technology and Language Revitalization: A Roadmap for the Mvskoke Language. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 7–12, St. Julians, Malta. Association for Computational Linguistics.

Stacey Mauger. 2025. The role of digital technology in Indigenous language revitalization: a systematic review of barriers, opportunities, and effective practices. Master's thesis, Ontario Tech University.

Paul J. Meighan. 2021. Decolonizing the digital landscape: the role of technology in Indigenous language revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 17.

Paul J. Meighan. 2024. Indigenous language revitalization using TEK-nology: how can traditional ecological knowledge (TEK) and technology support intergenerational language transmission? *Journal of Multilingual and Multicultural Development*, 45:3059–3077.

Marianne Mithun. 2001. *The languages of native North America*. Cambridge University Press.

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah Moeller, Jeffrey Micher, and 5 others. 2020. A Summary of the First Workshop on Language Technology for Language Documentation and Revitalization. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 342–351, Marseille, France. European Language Resources association.

Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu, and David Yarowsky. 2020. Fine-grained Morphosyntactic Analysis and Generation Tools for More Than One Thousand Languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3963–3972, Marseille, France. European Language Resources Association.

Alexandra O'Neil, Daniel Swanson, and Shobhana Chelliah. 2024. Computational Language Documentation: Designing a Modular Annotation and Data Management Tool for Cross-cultural Applicability. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 107–116, Bangkok, Thailand. Association for Computational Linguistics.

Flammie A Pirinen, Sjur N. Moshagen, and Katri Hiovain-Asikainen. 2023. GiellaLT — a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.

Robert Pugh and Francis Tyers. 2023. A finite-state morphological analyser for Highland Puebla Nahuatl. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 103–108, Toronto, Canada. Association for Computational Linguistics.

Mark Richards, Caroline Jones, Josephine Lardy, Anna Godden, Wanirr Godden, Sarah Bock, and Helena Lardy. 2025. Warrma Mangarrayi: Co-Designing an App for Learning Mangarrayi, an Indigenous Language of Northern Australia. *International Journal of Human–Computer Interaction*, 41(11):7172–7189.

Joshua Schwab-Cartas. 2018. Keeping Up with the Sun: Revitalizing Isthmus Zapotec and Ancestral Practices through Cellphilms. *The Canadian Modern Language Review*, 74(3):363–387.

Amalia Skilton, Sofia Gottlieb Pierson, Sunkulp Ananthanarayan, and Claire Bowern. 2025. Digital infrastructure and its impacts on language work: A case study of FieldWorks Language Explorer (FLEx). *Language*, 101(3):e136–e165.

Angela Taylor-Adams. 2019. Recording to revitalize: Language teachers and documentation design. *Language Documentation & Conservation*, 13:426–445.

Courtney Tennell and Kari AB Chew. 2024. Perspectives on relationality in online Indigenous language learning. *AlterNative: An International Journal of Indigenous Peoples*, 20(3):512–520.

Irina Wagner. 2017. New Technologies, Same Ideologies: Learning from Language Revitalization Online. *Language Documentation & Conservation*.

Joshua Wilbur. 2019. ELAN as a search engine for hierarchically structured, tagged corpora. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 90–103, Tartu, Estonia. Association for Computational Linguistics.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).